ANALYSIS OF MACROMOLECULAR STRUCTURE THROUGH EXPERIMENT AND COMPUTATION

A Thesis
Presented to
The Academic Faculty

by

John Jared Gossett

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Computational Science and Engineering

Georgia Institute of Technology
May 2013

ANALYSIS OF MACROMOLECULAR STRUCTURE THROUGH EXPERIMENT AND COMPUTATION

Approved by:

Dr. Stephen C. Harvey, Advisor
School of Biology
Georgia Institute of Technology

Dr. Alberto Apostolico
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. David Bader
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. Roger Wartell
School of Biology
Georgia Institute of Technology

Dr. Loren Williams
School of Chemistry and Biochemistry
Georgia Institute of Technology

Date Approved: 29 March 2013

ACKNOWLEDGEMENTS

I am indebted to the following people for their contributions to this dissertation:

To my advisor, Steve Harvey, for astutely guiding me through the process of obtaining a doctoral degree.

To my thesis committee—Alberto Apostolico, David Bader, Roger Wartell, and Loren Williams. Thank you for your valuable insight and criticisms.

To my collaborators, including Shreyas Athavale, Sarah Stabenfeldt, Chiaolong Hsiao, Jessica Bowman, Tim Lenz, and Lively Lie.

To past and present members of the Harvey Lab, including Anton Petrov, Burak Boz, Amanda McCook, Andrew Huang, Yingying Zeng, Bee Preeprem, Minmin Pan, Kanika Arora, Scott Douglas, Shefaet Rahman, Piyush Ranjan, Gaurav Sureka, and Ethan Speir, for being great colleagues.

And finally, to my parents, brothers, and friends. To all of you, from the bottom of my heart, thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS OR ABBREVIATIONS

BzCN	benzoyl cyanide.
MD	molecular dynamics.
MLD	maximum ladder distance.
NMIA	N-methylisatoic anhydride.
RMSD	root-mean-square deviation.
rRNA	ribosomal RNA.
SHAPE	2'-hydroxyl acylation analyzed by primer extension.
SPR	surface plasmon resonance.
STMV	satellite tobacco mosaic virus.

SUMMARY

This thesis covers a wide variety of projects within the domain of computational structural biology. Structural biology is concerned with the molecular structure of proteins and nucleic acids, and the relationship between structure and biological function.

We used molecular modeling and simulation, a purely computational approach, to study DNA-linked molecular nanowires. We developed a computational tool that allows potential designs to be screened for viability, and then we used molecular dynamics (MD) simulations to test their stability. As an example of using molecular modeling to create experimentally testable hypotheses, we were able to suggest a new design based on pyrrylene vinylene monomers.

In another project, we combined experiments and molecular modeling to gain insight into factors that influence the kinetic binding dynamics of fibrin "knob" peptides and complementary "holes." Molecular dynamics simulations provided helpful information about potential peptide structural conformations and intrachain interactions that may influence binding properties.

The remaining projects discussed in this thesis all deal with RNA structure. The underlying approach for these studies is a recently developed chemical probing technology called 2'-hydroxyl acylation analyzed by primer extension (SHAPE). One study focuses on ribosomal RNA, specifically the 23S rRNA from *T. thermophilus*. We used SHAPE experiments to show that Domain III of the *T. thermophilus* 23S rRNA is an independently folding domain. This first required the development of our own data processing program for generating quantitative and interpretable data from our SHAPE experiments, due to limitations of existing programs and modifications to the experimental protocol. In another study, we used SHAPE chemistry to study the *in vitro* transcript of the RNA genome of satellite tobacco mosaic virus (STMV). This involved incorporating the SHAPE data into

a secondary structure prediction program. The SHAPE-directed secondary structure of the STMV RNA was highly extended and considerably different from that proposed for the RNA in the intact virion.

Finally, analyzing SHAPE data requires navigating a complex data processing pipeline. We review some of the various ways of running a SHAPE experiment, and how this affects the approach to data analysis.

CHAPTER I

INTRODUCTION

1.1 Perspectives and Overview

Structural biology is concerned with the molecular structure of proteins and nucleic acids, and the relationship between structure and biological function. Computational techniques have revolutionized molecular modeling [76], which has become a sophisticated tool for investigating structure-function relationships. The projects discussed in this thesis demonstrate how experiments and computation can be used to analyze macromolecular structure.

1.2 Background

First, we define some key concepts used throughout the thesis. This is not intended to be a comprehensive review of the material. Where appropriate, I will direct the reader to articles and/or books that provide more detailed information.

1.2.1 Macromolecular modeling[1]

"Molecular modeling," writes Schlick, "is the science and art of studying molecular structure and function through model building and computation" [126]. Modeling on a computer—which, at a minimum, involves specifying the (x,y,z) coordinates of each atom—goes beyond just molecular graphics. We build models; we refine, optimize, and simulate them, studying their behavior over time. Most of the calculations involved would not be possible without the use of computers [76].

Molecular modeling finds many uses. It is essential for solving some experimental problems, such as structure refinement in X-ray crystallography. Modeling can also be used to supplement experimental approaches, such as the interpretation of structural data from an NMR experiment, or to solve problems that can't be solved experimentally, such as de novo structure prediction. Furthermore, we can use models pedagogically, to explain what

[1]Some parts of this subsection were adapted from lecture slides of Steve Harvey.

is already known about a protein or nucleic acid, or predictively, to develop experimentally testable hypotheses.

Molecular modeling is an important tool for understanding the relationship between structure and biological function. The structure-function relationship is a key concept in biology. Consider how genetic information is passed from one generation to the next. The Watson-Crick model for double-helical DNA provides the answer: with complementary base pairing (A–T and G–C), each strand serves as a template for the generation of the complementary strand [148]. Ultimately, it is the understanding of structure-function relationships that allows us to intervene in biological processes.

A popular computational approach to molecular modeling is molecular mechanics. Molecular mechanics involves a classical mechanical approximation in which atoms are point masses, bonds are springs, and quantum effects are ignored. Critical to the usefulness of molecular mechanics is calculating the internal energy of a given molecular conformation as accurately as possible. In reality there is a trade-off between accuracy and computational cost. The energy function (or force field) commonly includes terms that account for the deviation of bond lengths and bond angles from their equilibrium values, and for the rotation of bonds (torsions). Also included in the energy function are terms that describe interactions between non-bonded parts of the system, including van der Waal's and electrostatic interactions [76]. The calculation of non-bonded forces between pairs of atoms is a well-known bottleneck in molecular mechanics [126]. For a system of N atoms, the time needed to evaluate the forces scales as $O(N^2)$.

In addition to the energy function, the typical molecular mechanics computer program requires an iterative algorithm for generating successive conformations of the molecule. Molecular mechanics algorithms include energy minimization, the Monte Carlo method, and molecular dynamics (MD). Molecular dynamics is important for studying the structural and dynamic properties of molecular systems. Information such as molecular geometries and energies, rates of conformational changes, and protein folding pathways can be obtained from MD simulations [126]. Molecular dynamics represents the numerical integration of

Newton's equations of motion:

$$m_\alpha \ddot{\vec{r}}_\alpha = -\frac{\partial}{\partial \vec{r}_\alpha} U_{total}(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N), \quad \alpha = 1, 2, \ldots, N, \tag{1}$$

where m_α is the mass of atom α, $\vec{r}_\alpha$ is its position, and U_{total} is the total potential energy [115]. A popular program for running molecular dynamics simulations is NAMD [115].

1.2.2 RNA secondary structure prediction

The secondary structure of an RNA molecule is the collection of base pairs that are formed when it folds into a particular conformation. When referring to the secondary structure, we usually mean the secondary structure of the native conformation, unless explicitly stated otherwise. Predicting the secondary structure from the primary structure, or nucleotide sequence, is easier than predicting the three-dimensional structure, with which we are ultimately concerned, but it is a challenging problem nonetheless.

According to the thermodynamic hypothesis, the secondary structure with the lowest free energy is the predicted structure. In this case, one simply needs to estimate the free energy of folding for all possible secondary structures, and then select the one with the lowest free energy. This is actually a difficult problem because of the vast number of possible secondary structures: for an RNA of length N, the estimated number of possible secondary structure is $\sim 1.8^N$ [173]. This rules out a "brute force" approach [86]. Fortunately, dynamic programming algorithms, along with some simplifying assumptions, reduce the complexity to $O(N^3)$. Common programs for thermodynamic predictions of RNA secondary structure include mfold [172] and RNAstructure [89]. Also, GTfold is a parallelized secondary structure prediction program with significant improvements in runtime [138]. Unfortunately, dynamic programming algorithms like these do not always provide accurate secondary structure predictions.

Significant improvements to RNA secondary structure prediction accuracy can be achieved by applying folding constraints determined by chemical modification experiments [89]. One recent advance in using experimental data to improve secondary structure prediction is the incorporation of experimental SHAPE information as a pseudo-free energy

change term into a dynamic programming algorithm [38]. SHAPE, or 2'-hydroxyl acylation analyzed by primer extension, is a chemical probing technique designed to "report the extent to which a nucleotide is constrained by base pairing or other interactions" [93, 38].

Other approaches to secondary structure prediction include comparative sequence analysis, also known as covariation analysis, and knowledge-based methods. Comparative sequence analysis is appropriate only in cases when your RNA of interest has multiple divergent sequences with a common secondary structure [86].

1.3 Scope of Thesis

This thesis covers a wide variety of projects. Chapters 2 through 5 were adapted from publications that have appeared previously in peer-reviewed scientific journals [136, 53, 4, 3]. Since I was not the lead author on all of these publications, I feel compelled to clarify my contributions to each of them, which I do in the paragraphs below. I am also a co-author on two other peer-reviewed publications [5, 59], but these are not included in my thesis. Also not included here are two projects I worked on that have not been published. The first was a series of improvements I made to our in-house, Java-based molecular dynamics program, Oscar. (Oscar is primarily a teaching tool; it was not used for any of the work in this thesis.) The other unpublished project was a parallel algorithm I developed for converting from torsion space to Cartesian space.

In Chapter 2, using both experimental and modeling approaches, we investigate the interactions between fibrin "knob" peptides and fibrin "holes" to better understand fibrin assembly, an important mechanism for blood clot formation. This study provides insights for the rational design of knob mimics that more efficiently compete for hole occupancy. I designed, set up, and ran the molecular dynamics simulations, and I analyzed the simulation data. I also performed the electrostatic calculations.

Chapter 3 contains our modeling study on DNA-linked molecular nanowires. Molecular nanowires are composed of repeating molecular units designed to conduct electrical current. One strategy for creating a molecular nanowire from a DNA oligomer is to covalently link monomers to the DNA at regularly spaced positions, and then to chemically convert these

monomers into a conductive polymer. We use a molecular modeling approach to screen and evaluate potential DNA-linked polymer designs. I was the lead author on the paper from which this chapter was adapted.

We examine the ribosome in Chapter 4. Specifically, we test the hypothesis that Domain III is an independently folding domain of the 23S ribosomal RNA. We also discuss the evolutionary aspects of our results. This work involved comparing SHAPE experiments on Domain III in the intact 23S rRNA with Domain III as a separate RNA fragment. Due to limitations in existing SHAPE data processing tools I developed our own in-house MATLAB scripts to process the SHAPE data.

Viral RNA, specifically the RNA genome of satellite tobacco mosaic virus (STMV), is the subject of Chapter 5. Y. Zeng recently built an all-atom model of STMV [168]—believed to be the first all-atom model of any virus. This model was built using the experimentally and computationally determined secondary structure of Schroeder et al. [128]. In Chapter 5, we use SHAPE chemistry to probe the in vitro transcribed STMV RNA, and we compare our results with those from Schroeder et al. I processed the SHAPE data, performed the secondary structure predictions, and carried out the maximum ladder distance (MLD) calculations and analysis.

Chapter 6 discusses some aspects of the data analysis methodology required to interpret data from SHAPE experiments. This work has not been published.

I conclude the thesis in Chapter 7. I also offer my recommendations for future work.

CHAPTER II

BUILDING BETTER FIBRIN KNOB MIMICS: AN INVESTIGATION OF SYNTHETIC FIBRIN KNOB PEPTIDE STRUCTURES IN SOLUTION AND THEIR DYNAMIC BINDING WITH FIBRINOGEN/FIBRIN HOLES[1]

2.1 Abstract

Fibrin polymerizes via noncovalent and dynamic association of thrombin-exposed "knobs" with complementary "holes." Synthetic knob peptides have received significant interest as a means for understanding fibrin assembly mechanisms and inhibiting fibrin polymerization. Nevertheless, the inability to crystallize short peptides significantly limits our understanding of knob peptide structural features that regulate dynamic knob:hole interactions. In this study, we used molecular simulations to generate the first predicted structure(s) of synthetic knobs in solution before fibrin hole engagement. Combining surface plasmon resonance (SPR), we explored the role of structural and electrostatic properties of knob "A" mimics in regulating knob:hole binding kinetics. SPR results showed that association rates were most profoundly affected by the presence of both additional prolines as well as charged residues in the sixth to seventh positions. Importantly, analyzing the structural dynamics of the peptides through simulation indicated that the 3Arg side chain orientation and peptide backbone stability each contribute significantly to functional binding. These findings provide insights into early fibrin protofibril assembly dynamics as well as establishing essential design parameters for high-affinity knob mimics that more efficiently compete for hole occupancy, parameters realized here through a novel knob mimic displaying a 10-fold higher association rate than current mimics.

[1]This research was originally published in *Blood*. Stabenfeldt, S. E., Gossett, J. J., and Barker, T. H., "Building better fibrin knob mimics: an investigation of synthetic fibrin knob peptide structures in solution and their dynamic binding with fibrinogen/fibrin holes," *Blood*, vol. 116, no. 8, pp. 1352–1359, 2010.

2.2 Introduction

The activation and polymerization of the blood-circulating protein fibrinogen, a 340-kDa glycoprotein with 6 polypeptide chains $(A\alpha B\beta\gamma)_2$, is the primary homeostatic mechanism preventing excessive blood loss after vascular injury. This process is initiated by the activated serine protease thrombin, which specifically cleaves 4 N-terminal arginyl-glycine motifs on the 2 adjacent Aα and Bβ chains of fibrinogen, releasing 2 sets of fibrinopeptides A and B (FpA and FpB) and exposing cryptic fibrin polymerization knobs "A" and "B," respectively [13, 8, 85, 84, 75]. The newly exposed fibrin knobs noncovalently interact with complementary "holes" within the 2 distal C-terminal regions of the γ and β chains (complementary holes "a" and "b," respectively) to initiate fibrin protofibril assembly. Understanding the fundamentals of this dynamic and noncovalent knob:hole interaction will lead to both a more thorough understanding of fibrin assembly mechanisms and the establishment of design criteria for superior anticoagulants with high polymerization hole affinity to inhibit fibrin assembly.

Evidence for fibrin knob:hole interactions was first conclusively shown when fibrin polymerization was inhibited by synthetic knob A tripeptides (Gly-Pro-Arg) competing for fibrin holes [73, 72]. Characterization of the equilibrium binding affinities of both knob A and B peptide variants to fibrinogen showed that the knob A peptides (ie, GPRV and GPRP) have higher affinities to fibrinogen than knob B peptides (ie, GHRP and AHRP) under calcium-free conditions [73, 72, 74]. In the presence of calcium, the binding affinity knob B mimic GHRP significantly increases to near GPRP; however, GHRP is readily displaced by the knob A mimic GPRP, suggesting that knob A interactions are stronger than knob B [73]. Further evaluation of knob B peptide variants (ie, GHRPY, AHRPY, and MHRPY) showed the promiscuity of hole b versus hole a because the nonglycyl knob B peptides engaged hole b, but not hole a; only N-terminal glycyl peptides bind to hole a [40, 41]. Recent studies elegantly conducted with fragments from fibrinogen mutants and laser tweezers-based force spectroscopy further characterized A:a, A:b, and B:b interactions with native fibrin fragments [82, 81]. Consistent with the knob peptide studies, knob A interactions seem to dominate the knob:hole interactions because the A:a interaction displays a 6- to 8-fold

higher rupture force than A:b or B:b interactions [82, 81]. Although such steady-state, equilibrium studies have laid the foundation for understanding knob:hole interactions, investigating these binding events under dynamic conditions will provide critical information about the residence time of the noncovalent knob:hole interaction, a key determinant in fibrin assembly initiation and polymerization. In addition, understanding the fundamental structural cues within strong binding fibrin knob A mimics that drive the initial docking events and potentially stabilize the interaction (eg, enhance knob residence time within holes) will further establish design criteria required to develop superior anticoagulants that compete for hole occupancy.

Examination of the crystal structure of fibrinogen/fibrin hole regions (D fragment) with associated knob A peptides clearly established electrostatic interactions and hydrogen bonding between engaged knob peptides and holes a and b [134, 14, 18]. Crystal structures of D fragment generated with either GPRP or GPRVVE knob A peptides indicate that the 1Gly and 3Arg residues engage the same residues on the γ chain with minimal structural differences between the GPRP and GPRVVE [134, 14]. This observation leads one to question why GPRP displays a 4-fold greater affinity (K_D) for D fragment than for GPRV [73, 74]. Laudano and Doolittle [73] speculated that the higher affinity of GPRP was due to the 4-Pro residue potentially stabilizing the backbone of the GPRP, thereby reducing the degrees of freedom and the number of potential conformations. However, the structural properties of the knob peptides in aqueous environments before hole engagement have not been explicitly examined largely because of the inability to crystallize small peptides for structural x-ray studies. Such knowledge is critical for rationally designing knob peptides, as well as fully synthetic analogs, with superior anticoagulant properties than is currently available (ie, GPRP). Molecular modeling and molecular dynamics (MD) simulations are an emerging approach that explores the conformational landscape of short peptides enabling one to assess molecular structural differences that influence functional binding parameters [80, 57].

In this study, we investigated fibrin knob peptide:hole interactions with both experimental and theoretical modeling approaches to elucidate factors that influence the kinetic

binding dynamics (k_a and k_d) of knob A peptide variants to fibrin holes. We focused on (1) kinetic modeling of the binding interaction and (2) structural characterization of the peptides in solution. Previous binding affinity studies have contributed significantly to the current understanding of knob:hole interactions; however, as described earlier, these seminal experiments were performed under equilibrium conditions in which the details of critical dynamic interactions are overlooked. Using surface plasmon resonance (SPR), we evaluated the kinetic binding interactions of fibrin knob peptides with fibrinogen/fibrin holes and investigated appropriate kinetic binding models to describe the knob:hole interaction. On the basis of past literature, we examined a set of knob A peptide variants of 7 to 8 residues in length to evaluate 2 properties hypothesized to influence binding kinetics (Table 1). The first property included sequences with backbone "stabilizing" residue configurations such as a single Pro [90, 125] or Pro-Pro [141]. Second, we chose residues that would alter the charge distribution across the chain similar to those observed on the native knob A chain (ie, arginine and glutamic acid). Subsequent molecular modeling and dynamic simulations of each peptide facilitated structural comparisons of the peptide conformations in solution. In this report, we correlate molecular/structural properties of the knob peptide residues with functional kinetic binding parameters to gain a better understanding of knob characteristics that contribute to knob:hole interactions and identified potential criteria for the rational design of enhanced knob variants. Illustrating this point, we identified and report here a novel knob peptide mimic with a unique element (Pro-Phe-Pro) that enhances the association rate to polymerization holes nearly 1 order of magnitude.

2.3 Methods

2.3.1 Fibrin knob A peptides

Peptide sequences included GPRVVAAC, GPRVVERC, GPRPAAC, GPRPPERC, GPRPFPAC, and GPSPAAC (GenScript Inc; Table 1). The peptide sequences were designed with a carboxyl-terminal cysteine residue to permit sulfhydryl-targeted reactions for future conjugation chemistries.

Table 1: Experimental knob A peptides and corresponding properties.

Peptide sequence	Property	Net charge
GPRVVERC	Mimics native sequence through seventh residue	+1
GPRVVAAC	Mimics native sequence minus additional charged residues	+1
GPRPAAC	Stabilized backbone	+1
GPRPFPAC	Stabilized backbone	+1
GPRPPERC	Stabilized backbone and additional charged residues	+1
GPSPAAC	Negative control; known dysfibrinogen mutant	0

2.3.2 Preparation of fibrinogen D fragment

Human fibrinogen (Enzyme Research Laboratories) at 2 mg/mL was digested with 0.1 U/mL human plasmin (Enzyme Research Laboratories) in HEPES (N-2-hydroxyethylpiperazine-N'-2-ethanesulfonic acid) + $CaCl_2$ buffer (150mM NaCl, 5mM $CaCl_2$, 25mM HEPES; pH 7.4) overnight at room temperature. D fragment was isolated as previously described, with slight modifications [68]. Briefly, the plasmin-digested fibrinogen and GPRPAA beads were incubated for 30 minutes, with occasional agitation. The unbound proteins and protein fragments were removed with excessive washing with HEPES + $CaCl_2$ buffer. D fragment was eluted with 1M sodium bromide and 50mM sodium acetate (pH 5.3). Eluted samples were pooled together and exchanged back into HEPES + $CaCl_2$ buffer with a centrifugal filter (molecular weight cutoff, 10 000 Da). D fragment was verified by sodium dodecyl sulfate-polyacrylamide gel electrophoresis and stored at $-80°C$ until use.

2.3.3 Binding kinetics with SPR

The Biacore 2000 (Biacore Lifesciences, GE Healthcare) was used investigate kinetic binding constants (k_a and k_d) of knob A peptide variants for fibrinogen D fragment. Briefly, D fragment was covalently immobilized to gold-coated SPR sensor chips via self-assembled

monolayer surface chemistry to generate a nonfouling surface with a controlled density of reactive carboxylic acid groups. Mixed self-assembled monolayers were generated on gold-coated chips as described previously [113] by incubating with a 1-mM mixture of tri(ethylene glycol)-terminated alkanethiols (HS–$(CH_2)_{11}$–$(OCH_2CH_2)_3$–OH; ProChimia) and carboxylic acid-terminated alkanethiols (HS–$(CH_2)_{11}$–$(OCH_2CH_2)_6$–OCH_2COOH) for 4 hours. On loading the senor chip into the Biacore 2000, the carboxylic acid-terminated alkanethiol in all 4 flow cells was activated by flowing 200mM 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (Sigma-Aldrich) and 50mM N-hydroxysuccinimide (Sigma-Aldrich; 5 μL/minute for 10 minutes). Immediately after activation, D fragment was immobilized in 3 flow cells (5 μL/minute for 10 minutes) to achieve 1800 to 2000 resonance units (1 resonance unit $\sim$ 1 pg/mm^2). Unreacted N-hydroxysuccinimide groups were quenched in all 4 flow cells (3 sample cells and 1 reference cell) with 20mM ethanolamine (10 μL/minute for 10 minutes). On stabilization of the baseline signal, kinetic binding experiments were run in duplicate with the peptide variants as the flow analytes. Five various concentrations for each peptide (0.94 μM to 150 μM) were flowed at 25 μL/minute for 4 minutes immediately followed by a 10-minute dissociation phase. Between each run, the surface was regenerated with 1M sodium bromide and 50mM sodium acetate (pH 6.0). SPR experiments were performed 3 times with varying peptide injection order to rule out binding trends associated with injection sequence. Peptide solutions were incubated with tris(2-carboxyethyl)phosphine immobilized on agarose beads (Thermo Fisher Scientific Inc) to ensure reduction of any disulfide bonds between C-terminal cysteines. Mass spectrometry analysis (Fast Atom Bombardment) of peptide solutions showed that the peptides did not dimerize over the course of the SPR experiment (Figure 25, Appendix A).

2.3.4 SPR analysis and evaluation

SPR sensorgrams were analyzed with the aid of Scrubber 2 and ClampXP software (Center for Biomolecular Interactions Analysis, University of Utah) [100, 101, 102]. Before analysis, all sensorgrams were inspected for abnormalities (ie, baseline drift, air spikes, or irregular

deviations) and excluded. Reference cell responses were subtracted from corresponding active response curves. Double-referenced curves were acquired by further subtracting the reference cell blank buffer injections from each reference-subtracted response curve [103]. All double-referenced curves were normalized by the molecular weight of each peptide and multiplied by 1000 to account for minor variations in response because of molecular weight. The resulting curves were then analyzed and fitted to the kinetic models. Kinetic modeling and simulations were performed with ClampXP software with the Langmuir 1:1 model or the heterogeneous ligand model; globally fitted parameters were determined for each kinetic dataset per peptide. Equilibrium binding constants were calculated from fitted kinetic constants. Goodness of fit for each model was determined by evaluating the residual plots and residual sum of squares [103].

2.3.5 MD simulations

Classical MD simulations were performed with 5 knob A peptides, GPRVVAAC, GPRVVERC, GPRPAAC, GPRPPERC, and GPRPFPAC. Because the crystal structure of each of these peptides within the fibrin hole has not been determined experimentally, the initial peptide structures were rendered in Swiss-PDB Viewer (Swiss Institute of Bioinformatics [54]) with the backbone torsion angles of the first 3 residues constrained to an "active" peptide conformation obtained from previously published D fragment crystal structures (PDB code: 2HPC and 2FFD [14]). Before MD simulations, the structure of each peptide was minimized with 10 iterations of steepest descent (500 steps) energy minimization in vacuo. Each peptide was placed in the center of a water box (Visual Molecular Dynamics software [62]) supplemented with Na^+ and Cl^- ions to achieve electric neutrality, mimicking experimental conditions ($\sim$340 mOsmol/L). The models were initially minimized for 1000 steps with the backbone atoms fixed, followed by 1000 steps of minimization with harmonic restraints on the α carbon atoms. After energy minimization, each system was heated to 310K over a period of 20 picoseconds with harmonic restraints on the α carbons. Next, with the restraints still active, each system was equilibrated at constant temperature (310 K) and pressure (1 atm) for 100 picoseconds. The restraints were then removed, and

the equilibration was continued for 200 picoseconds. The production runs were carried out for 10 nanoseconds under constant temperature and pressure conditions, ie the *NPT* ensemble. Temperature was maintained at 310 K, pressure at 1 atm. Short-range nonbonded interactions were cut off at a distance of 1.2 nm (12 Å) with a switching function between 1.0 and 1.2 nm (10 and 12 Å). The particle mesh Ewald method was used to compute electrostatics [35]. All bonds involving hydrogen atoms were constrained with the use of the SHAKE algorithm [121], which allowed for an integration time step of 2 femtoseconds. All simulations were performed with NAMD Version 2.6 (Theoretical and Computational Biophysics Group at University of Illinois at Urbana-Champaign [115]) with the use of the CHARMM22 force field parameter set [88].

2.3.6 MD simulation analysis

Clustering analysis. A hierarchical cluster analysis was performed on the trajectory data from each peptide MD simulation. A trajectory for clustering was obtained by taking every 100th frame (100-picosecond interval) from a 10-nanosecond production run. Next, we calculated the root mean squared deviation (RMSD) between every frame in the trajectory to generate a dissimilarity matrix. RMSD was calculated on the basis of the backbone atoms after optimal superposition. The *agnes* function in the *cluster* package supplied with the R statistical software package (R Project) was used to construct a hierarchy of clusterings from the dissimilarity matrix. The clusters were visualized in VMD with the use of the Cluster plugin. On the basis of the resulting dendrograms generated from the cluster analysis (Figure 28, Appendix A), representative trajectory conformations from the 2 most populated cluster groupings at the third level were used to compare both conformational and electrostatic properties between each peptide.

Structure/conformation and electrostatic properties comparisons. For structural/conformational comparison, representative conformations from the top 2 populated clusters were superimposed onto either GPRP or GPRV peptides in an active conformation within hole a as obtained from previous published D fragment crystal structures (PDB code: 2HPC and 2FFD [14], respectively); GPRPxxx peptides were compared with active GPRP

and GPRVxxx peptides were compared with active GPRV. After least-squares superpositioning the first 3 residues of each conformation along the backbone, the RMSD of the first 3 residues was calculated with the active GPRP or GPRV as the reference; both the backbone and total atom RMSDs were calculated. For electrostatic comparisons, electrostatic potential surface maps for representative conformations from the top populated cluster group for each peptide were generated with Adaptive Poisson-Boltzmann Solver with the use of the CHARMM22 force field parameter set [9].

2.4 Results

2.4.1 Kinetic binding models

To investigate the dynamic binding profile between the fibrinogen/fibrin holes and knob peptide variants, we used SPR. Binding interactions were evaluated by flowing the knob peptides over an immobilized surface of D fragment (Figure 1). By immobilizing D fragment as opposed to full-length fibrinogen, we simplified the kinetic binding model to a heterogeneous 2-site ligand model (ie, 1 hole a and hole b per ligand) as opposed to a 4-site model (ie, 2 of each hole a and hole b per ligand). We modeled the data with the use of both a Langmuir 1:1 model and a heterogeneous ligand model to compare previously established binding affinities to a more dynamic 2-site model. However, the complexity of the heterogeneous model fitted parameters for sites 1 and 2 (ie, maximal binding response, k_a, and k_d) limits direct designation or assignment of holes a or b to site 1 or 2. An additional mass transport limited model was tested as well (data not shown), but it did not fit the experimental data for any of the peptides. All peptide SPR data were fit except for the negative control peptide, GPSPAAC, in which minimal binding response was observed.

In comparing model simulation results, the heterogeneous ligand model fit the experimental binding data far better than the Langmuir 1:1 model. Here, we present response, simulation curves, and residual plots for a single set of experimental data (GPRPFPAC; Figure 2); plots for all 5 peptides are provided in Figures 26 and 27 (Appendix A). Looking specifically at the residual sum of squares, the range for the 1:1 Langmuir model (1.497–2.197; Table 2) was higher than the heterogeneous ligand model (0.9437–1.474; Table 3),

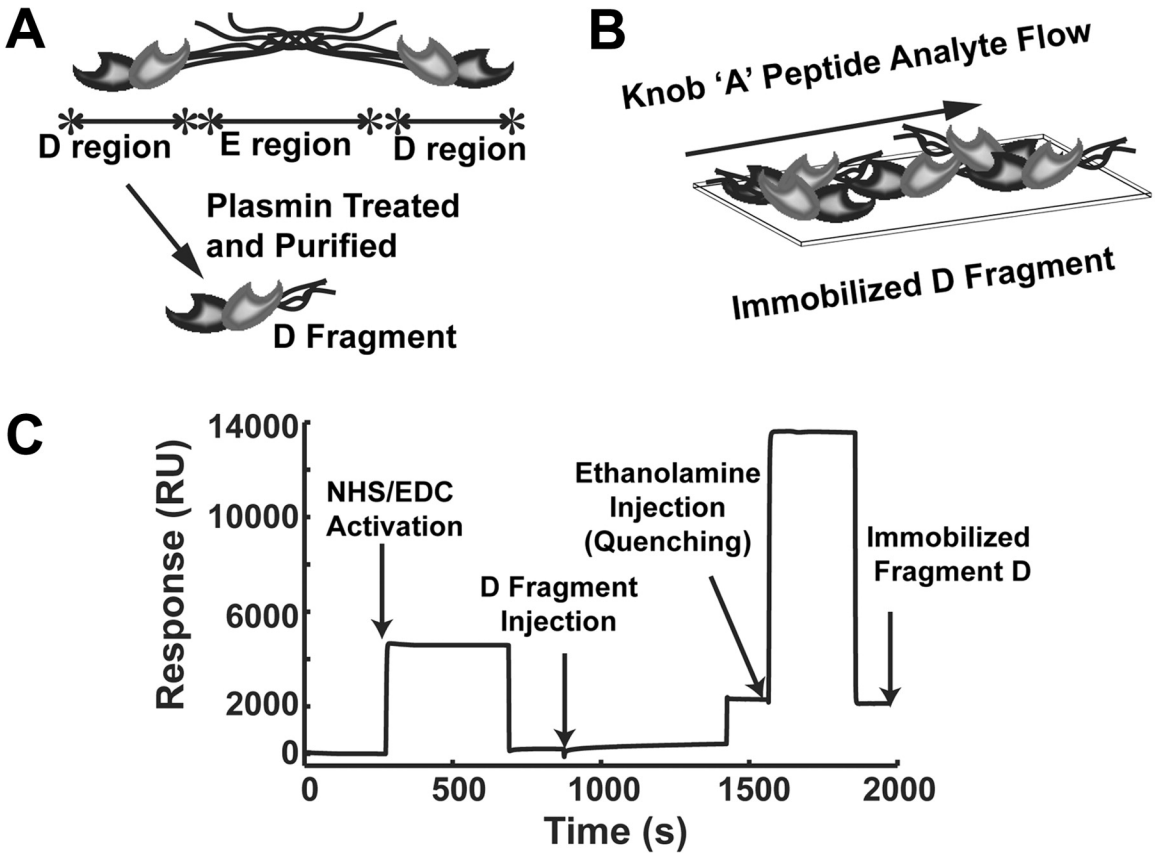


Figure 1: SPR experimental protocol. (A) Schematic representation of fibrinogen and the 2 major regions, E and D. Plasmin treated fibrinogen, and purification of the fragments generates D fragment. (B) Surface plasmon resonance (SPR) experimental set-up with D fragment immobilized to an SPR chip acting as the ligand and the knob A peptides flow across the surface as the analyte. (C) Representative SPR sensorgram for D fragment immobilization where the carboxyl-terminated self-assembled monolayers were activated by 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDC)/N-hydroxysuccinimide (NHS), enabling amine-targeted immobilization of D fragment. Ethanolamine quenched any unreacted carboxyl groups and rid the surface of nonspecifically bound D fragment.

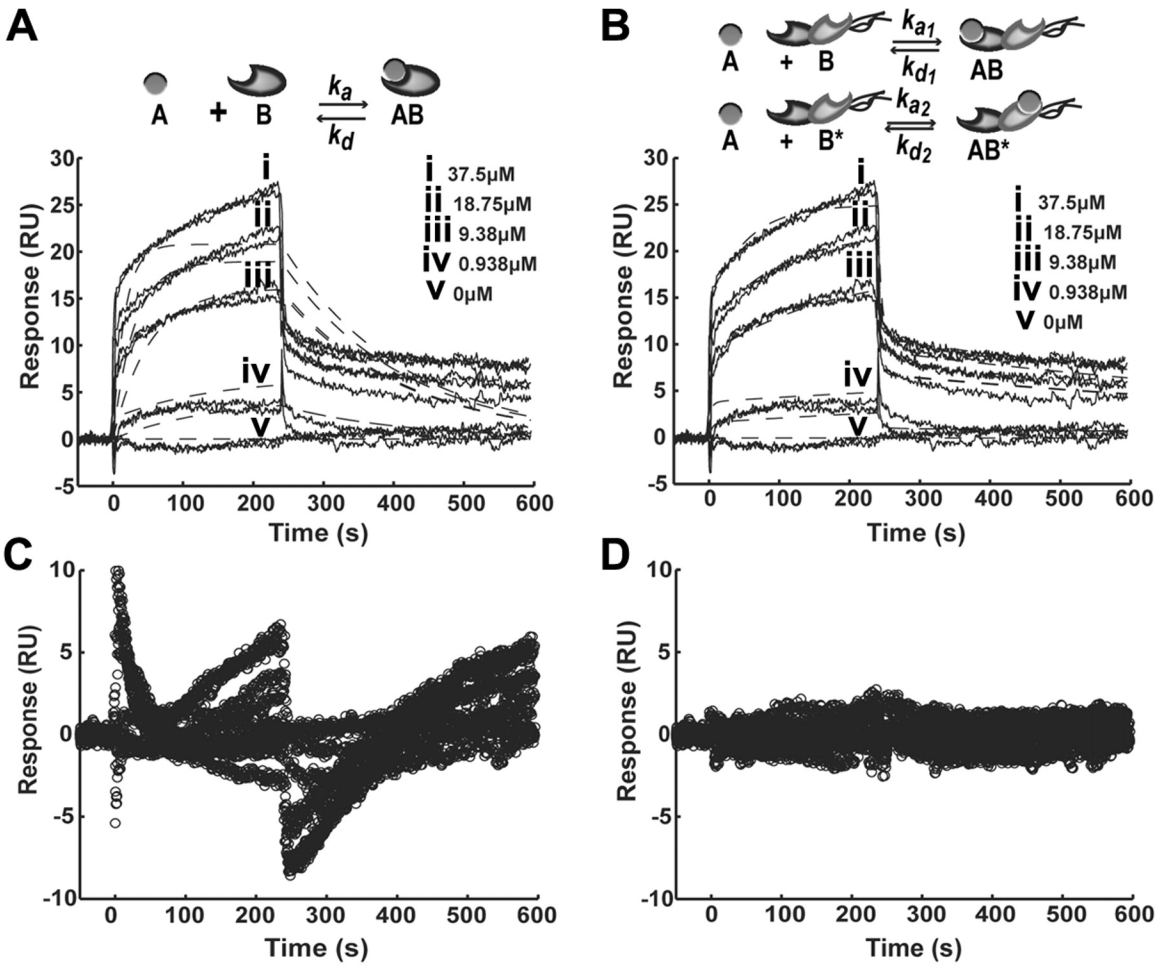


Figure 2: Kinetic model comparison. Experimental sensorgram of GPRPFPAC fitted with (A) Langmuir 1:1 model or (B) heterogeneous ligand model. Corresponding residuals plots for (C) Langmuir model or (D) heterogeneous ligand model. Solid lines indicate experimental SPR response curves; dashed lines, fitted model curves.

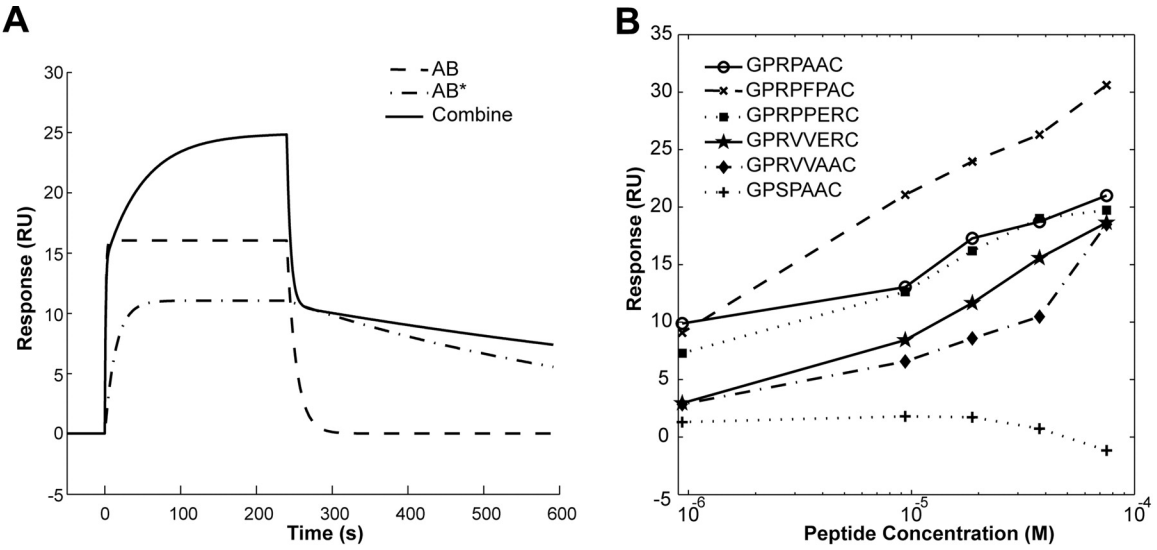


Figure 3: Contribution of AB and AB* binding in 2-site model and maximal binding response. (A) Simulation sensorgrams generated by the heterogeneous ligand model for GPRPFPAC. The combine response is a sum of analyte-ligand complexes AB and AB*. (B) Maximal binding response (resonance unit; RU) for the corresponding peptide concentration (molar, M). GPSPAAC, GPRPAAC, GPRPFPAC, GPRPPERC, GPRVVERC, and GPRVVAAC.

suggesting that the fitted heterogeneous ligand model deviated less from the experimental data. In addition, graphically plotting the residuals over time showed that the residuals for 1:1 Langmuir model followed a systematic trend (Figure 2C), indicative of fitting an inappropriate model to the experimental data [33]. In contrast, the residuals for the heterogeneous model were lower and more randomly distributed (Figure 2D), indicating that this model adequately describes the binding response curves [100, 33]. On the basis of these analyses and observations, further comparisons of binding parameters were performed with the results from the heterogeneous ligand model.

2.4.2 Fitted binding affinity parameters

The fitted parameters (B_{max}, k_a, and k_d) for each knob peptide variant for the heterogeneous ligand model are displayed in Table 3. In addition, the sensorgram plots in Figure 3A show the contribution each fitted parameter set has on the overall combined 2-site model. For example, for GPRPFPAC, the fit for the AB complex (site 1) encompassed a fast association rate, presumably accounting for the large initial response, whereas the slower

association rate for the AB* complex (site 2) contributes to the slower response for the duration of the injection. Broadly comparing all the knob peptide variant parameters, the 4Pro peptides (ie, GPRPAAC, GPRPFPAC, and GPRPPERC) had much faster association rates (k_{a1} = 2.84–21.72 $\times 10^3 M^{-1}s^{-1}$; k_{a2} = 1.01–1.81 $\times 10^3 M^{-1}s^{-1}$) than the 4Val peptides (ie, GPRVVAAC and GPRVVERC; k_{a1} = 0.62–1.07 $\times$ $10^3 M^{-1}s^{-1}$; k_{a2} = 0.04–0.26 $\times$ $10^3 M^{-1}s^{-1}$).

In comparing the 4Pro variants, one of the most striking differences was the nearly 10-fold increase in k_{a1} for GPRPFPAC (21.72 $\times 10^3 M^{-1}s^{-1}$) compared with GPRPAAC (2.84 $\times$ $10^3 M^{-1}s^{-1}$) and GPRPPERC (3.22 $\times$ $10^3 M^{-1}s^{-1}$); however, for k_{d1}, GPRPAAC (12.83 $\times$ $10^{-3}s^{-1}$) displayed a 6-fold slower rate compared with GPRPFPAC (81.10 $\times$ $10^{-3}s^{-1}$). In contrast, for the second binding site the k_{a2} rate for GPRPFPAC (1.81 $\times$ $10^3 M^{-1}s^{-1}$) was only moderately faster than GPRPAAC (1.05 $\times$ $10^3 M^{-1}s^{-1}$) and GPRPPERC (1.01 $\times$ $10^3 M^{-1}s^{-1}$), whereas the k_{d2} for GPRPFPAC (1.96 $\times$ $10^{-3}s^{-1}$) was nearly 8-fold slower than GPRPAAC (8.95 $\times 10^{-3}s^{-1}$). These simulation results indicate that GPRPFPAC has a higher affinity to the first and second binding sites and additionally dissociates more slowly from the second binding site, thus translating to longer engagement in fibrinogen holes compared with the other variants tested.

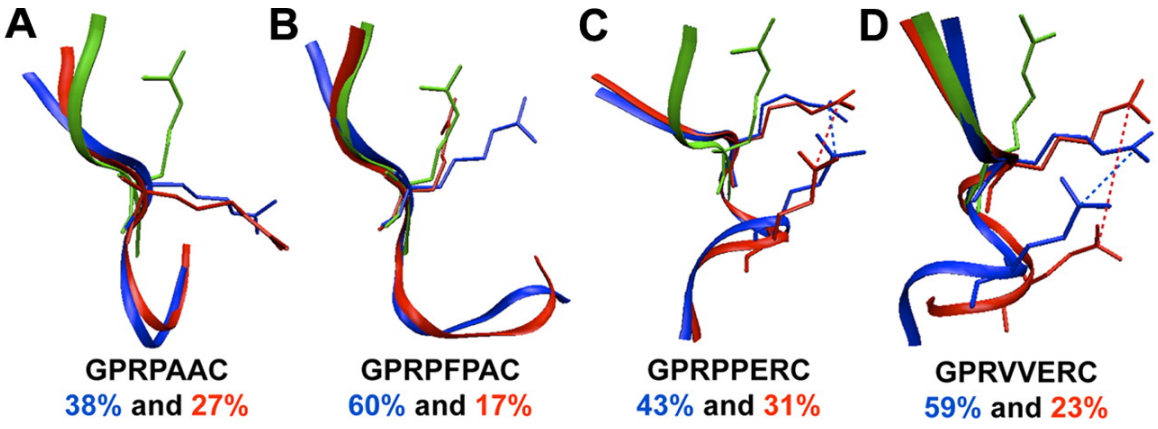
We also investigated the effect additional charged residues in the sixth and seventh position had on functional binding characteristics by comparing GPRVVAAC and GPRVVERC. For the association rates, GPRVVERC had a 2-fold increase over GPRVVAAC in k_{a1} (1.07 $\times$ $10^3 M^{-1}s^{-1}$ vs 0.62 $\times 10^3 M^{-1}s^{-1}$, respectively) and a 6-fold increase in k_{a2} (0.26 $\times 10^3 M^{-1}s^{-1}$ vs 0.04 $\times$ $10^3 M^{-1}s^{-1}$, respectively). However, the dissociation rates for GPRVVERC were 2-fold faster than GPRVVAAC for k_{d1} (57.67 $\times$ $10^{-3}s^{-1}$ vs 30.08 $\times$ $10^{-3}s^{-1}$, respectively) and 4-fold faster for k_{d2} (4.07 $\times$ $10^{-3}s^{-1}$ vs 1.00 $\times$ $10^{-3}s^{-1}$, respectively). These results collectively imply that, although the additional charged residues (ie, 6Glu and 7Arg) may enhance the affinity of the knob peptide to the binding holes, it may also result in an increased rate of dissociation.

2.4.3 Equilibrium dissociation constants

Using the fitted k_as and k_ds from the kinetic models, we calculated the equilibrium dissociation constants (K_Ds; Table 3). These results are also represented graphically by plotting the SPR binding maximum for each variant versus injection concentration (Figure 3B). The lowest K_Ds were observed in the peptide variants with a 4Pro (GPRPFPAC < GPRPAAC < GPRPPERC < GPRVVERC < GPRVVAAC). In comparing the specific K_Ds for each site between the 4Pro variants, K_{D1} for GPRPFPAC (3.73 μM) and GPRPAAC (4.53 μM) was significantly lower than GPRPPERC (18.23 μM). However, for the second site, the K_{D2} values for GPRPFPAC (1.08 μM) and GPRPPERC (2.93 μM) were significantly lower than GPRPAAC (8.52 μM). This result further indicates that GPRPFPAC interacts and engages the hole domains more readily than the other variants tested, even the gold standard GPRP mimic (GPRPAAC). In comparing GPRVVAAC with GPRVVERC, the addition of charged residues, 6Glu and 7Arg, decreased K_{D2} from 25.00 μM to 15.71 μM, whereas K_{D1} was relatively similar.

2.4.4 MD simulation analysis

MD simulations were performed to compare conformational structures of the peptides immersed in an aqueous environment before engagement with fibrin holes. Because of the large amount of data (ie, conformations) in a 10-nanosecond MD trajectory, a hierarchical clustering method was used to select a smaller set of conformations representative of the entire MD trajectory (Figure 28, Appendix A). From this cluster analysis, representative conformations from the 2 most populated clusters for each peptide were used to compare structural features between experimental peptides. The representative conformations for each peptide were then superimposed on either active GPRP or GPRV peptides obtained from D fragment crystal structures (PDB code: 2HPC and 2FFD [14], respectively; Figure 4). The superposition of simulation conformations with active conformation was evaluated by calculating the RMSD for atoms (backbone and all atoms) from the first 3 residues of the simulation conformations in reference to the known active conformations. The lower the RMSD, the better the alignment to the reference point. Ranking the peptides from lowest to



F

Peptide	Population (% of total)	Backbone RMSD (Å)*	Total RMSD (Å)*
GPRPAAC	38%	0.799	4.564
	27%	0.803	4.709
GPRPFPAC	60%	0.594	4.615
	17%	0.655	4.802
GPRPPERC	43%	0.862	4.896
	31%	0.623	4.894
GPRVVERC	59%	0.629	5.042
	23%	0.834	5.251
GPRVVAAC	48%	1.546	5.437
	28%	0.717	4.757

Figure 4: Structural analysis. Representative trajectories of the top 2 (blue and red) most populated groups from hierarchical cluster analysis superimposed on the active GPRP or GPRV conformation (green); GPRPxxx peptides were compared with active GPRP and GPRVxxx peptide were compared with active GPRV. (A) GPRPAAC, (B) GPRPFPAC, (C) GPRPPERC, (D) GPRVVERC, and (E) GPRVVAAC. (F) The population percentage represents the percentage of total number trajectories in 1 conformational cluster; the top 2 populated clusters at the fourth level of the hierarchical cluster are reported. RMSD calculations for the first 3 residues were in reference to the active conformation (ie, GPRP or GPRV); both backbone and total RMSDs were calculated after optimal superposition along the backbone.

highest weighted average RMSD along the backbone was GPRPFPAC (0.0607 nm [0.607 Å]) less than GPRVVERC (0.0670 nm [0.670 Å]) less than GPRPPERC (0.0761 nm [0.761 Å]) less than GPRPAAC (0.0801 nm [0.801 Å]) less than GPRVVAAC (0.1241 nm [1.241 Å]). It appears that the Pro-Phe-Pro residues in GPRPFPAC help stabilize the backbone of the first 3 residues to a conformation similar to active GPRP. Similarly, although GPRVVERC and GPRPPERC were chosen to investigate the alterations in electrostatic charge, a salt bridge developed between the 3Arg and 6Glu side chains potentially stabilizing the backbone. We also calculated the RMSD of all the atoms (ie, backbone and side chain atoms) in the first 3 residues; here, the weighted average RMSD ranking from lowest to highest was GPRPFPAC (0.4656 nm [4.656 Å]) less than GPRPAAC (0.4657 nm [4.657 Å]) less than GPRPPERC (0.4895 nm [4.895 Å]) less than GPRVVERC (0.5101 nm [5.101 Å]) less than GPRVVAAC (0.5186 nm [5.186 Å]). This RMSD ranking inversely correlated with the experimental binding affinity data (ie, lower RMSD, higher binding affinity). Considering this correlation, we evaluated the orientation of the side chain groups in comparison to the active conformation, particularly the orientation of 3Arg, which is required for binding of fibrin holes. Conventional terminology for torsional side chain angle defines χ_1 as the angle between N_i–$C_{\alpha i}$–$C_{\beta i}$–$C_{\gamma i}$ [28]; the 3 common rotamer classifications are gauche$^-$ (0° to 120°), trans (120° to 240°), and gauche$^+$ (−120° to 0°) [116]. In the active conformation for both GPRP and GPRV, the 3Arg χ_1 angle is in a gauche$^+$ conformation. However, in assessing the χ_1 angle of the 3Arg side chain for GPRPAAC and GPRVVAAC, we noted the angle was predominantly in a trans conformation during the simulation (ie, pointing toward the carboxyl-terminus of the sequence; Figure 4A,E). In contrast, the 3Arg group in GPRPFPAC was mobile, but predominantly in the gauche$^+$ conformation and rarely the trans conformation (Figure 4B). We speculate that the bulky side chain on 5Phe sterically hinders electrostatic interactions between 3Arg and the C-terminus as observed in GPRVVAAC and GPRPAAC. In addition, the salt bridge formation between 3Arg and 6Glu in GPRVVERC and GPRPPERC appeared to stabilize the 3Arg side chain in the gauche$^+$ conformation (Figure 4C–D). Collectively, these observations suggest that the 3Arg rotameric classifications depended on properties of the downstream residues.

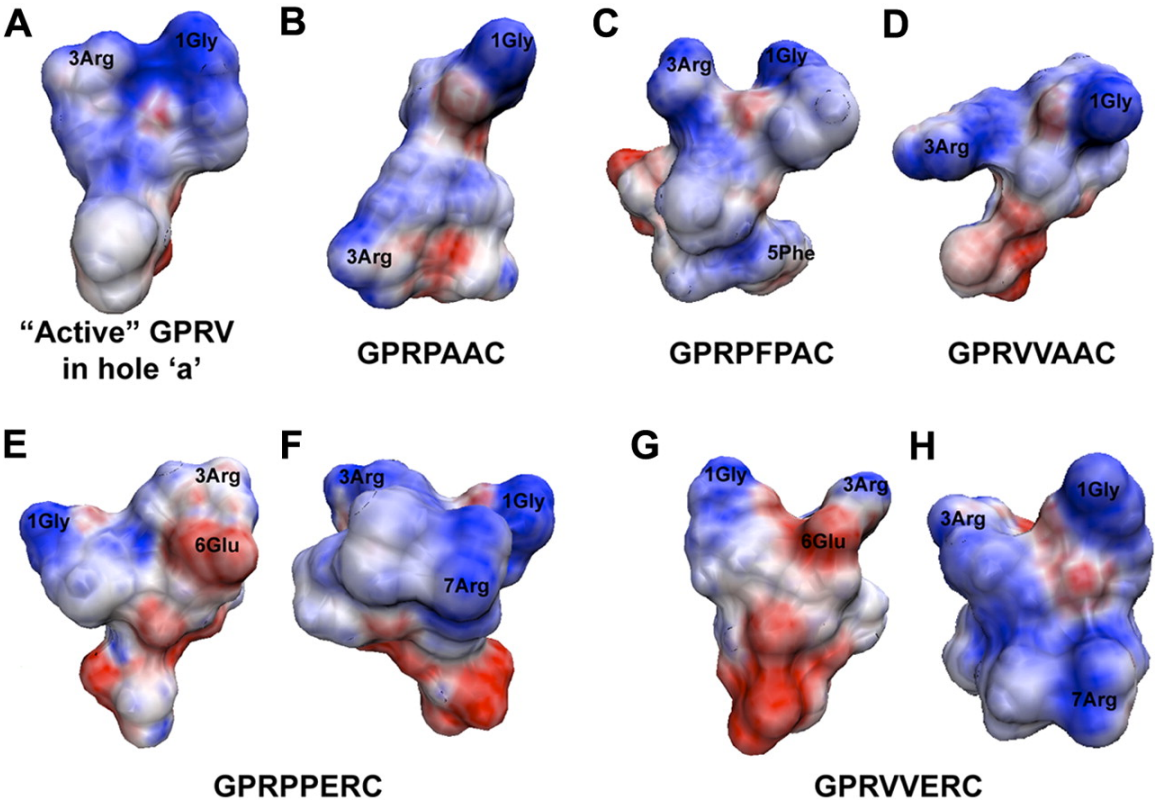


Figure 5: Electrostatic potential surface maps. Electrostatic potential surface maps were for the representative trajectory from most occupied cluster grouping. (A) GPRV in the active conformation, (B) GPRPAAC, (C) GPRPFPAC, (D) GPRVVAAC, (E) GPRPPERC, (F) GPRPPERC rotated 180° about the vertical axis, (G) GPRVVERC, and (H) GPRVVERC rotated 180° about the vertical axis.

As previously mentioned, knob:hole interactions are driven by electrostatic interactions. Therefore, we generated electrostatic potential surface maps to display the charge distributions for each peptide variant (Figure 5). The active conformation of GPRV has a noticeably positively charged N-terminus generated by 1Gly and 3Arg (Figure 5A; GPRP was not shown, but it has similar structure/map properties). For the peptide variants, we noted that the electrostatic mapping was directly related to the 3Arg side chain rotamer classification, in which the gauche$^+$ 3Arg conformations maintained the concentrated positive charge around 1Gly and 3Arg (ie, GPRPFPAC; Figure 5C). However, with 3Arg in the trans conformation (ie, GPRPAAC and GPRVVAAC), the positive charge was more broadly distributed from the N-terminus across to the C-terminus (Figure 5B,D). The addition of -ERC in the sixth through eighth residues resulted in 2 alterations (Figure 5E-H). First, although the salt bridge between 3Arg and 6Glu stabilized 3Arg in the gauche$^+$ conformation providing a concentrated positive charge at the N-terminus, the presence of 6Glu contributed a slightly negative charge near 3Arg (Figure 5E,G). Second, the additional Arg residue in the seventh position redistributed the positive charge more broadly across the peptide chain (Figure 5F,H). Collectively, these MD simulations and subsequent analyses provided snapshots at the molecular level into potential intrachain interactions that occur in aqueous environments and may contribute to the initial binding interactions with fibrin holes.

2.5 Discussion

In an effort to both describe, for the first time, fibrin knob structure in solution (ie, before complimentary hole engagement) and explore potential factors that affect the initial docking/binding of fibrin knobs to holes, we modeled the binding kinetics of fibrin knob peptide:hole interactions and investigated structural properties of the peptide variants in aqueous environments. Results from this study provide significant insights into the structural dynamics of knob sequences and the role of structure in defining the dynamics of knob:hole interactions that govern fibrin assembly and fiber structure. Furthermore, these studies enabled the discovery of a novel knob mimic that displays an association rate to

fibrin polymerization holes an order of magnitude higher than any previously published knob sequence.

Three decades have passed since Laudano and Doolittle [72] first reported that the short tripeptide (GPR) derived from knob A binds to fibrinogen. By further extending the peptide sequence to 4 residues, Doolittle noted that the synthetic tetrapeptide GPRP (20 μM) had a higher affinity to fibrinogen compared with the human knob A mimic GPRV (75 μM) and that both knob A variants bound to more than 3 sites on fibrinogen [73, 72, 74]. Recent x-ray crystallographic studies of D fragment in the presence of either GPRP or GPRVVE clearly verified that both knob A mimics were capable of occupying both holes a and b [14]. These studies also established the remarkably high degree of structural similarity between GPRP and GPRV in the engaged position, but they were unable to elucidate why GPRP displays significantly higher affinity for fibrin holes. Previous reports have speculated that the enhanced binding affinity of GPRP over GPRV may be due to restrictions the 4Pro imparts on the peptide backbone [73], yet this has been an unverified theory until now. We addressed this critical gap by characterizing the binding kinetics of a set of knob A variants designed to specifically investigate the effect additional backbone stabilizing residues (ie, Pro and Phe) and/or electrostatic charged residues (ie, 6Arg and 7Glu mimicking the native human knob A sequence) [16] have on the binding dynamics of knob:hole interactions. We were able to capture the more complex and dynamic interactions within each hole using a heterogeneous ligand model, a model that correlates with Doolittle's initial findings that A:b interactions, particularly knob A peptide:hole b interactions can and do occur [74, 14]. Furthermore, we discovered a novel peptide, GPRPFPAC, with the highest reported affinity to the hole domains, even surpassing the binding activity of the gold standard GPRP mimic (GPRPAAC).

Recent SPR studies have investigated the interaction between adsorbed fibrinogen and the N-terminal disulfide knot (NDSK) of differentially activated fibrin (FpA and FpB removed = desAB-NDSK; only FpA removed = desA-NDSK) [50]. The investigators reported K_Ds of 5.8 μM and 3.7 μM for desA-NDSK and desAB-NDSK, respectively. The slightly higher affinity of desAB-NDSK was attributed to B:b interactions because coinjection of

knob B peptides with desA-NDSK or desAB-NDSK hindered only the desAB-NDSK interaction with fibrinogen [50] Even though this elegant study established the presence of B:b binding interactions, these SPR experiments were performed under equilibrium conditions (ie, low flow rates) and determined only a single equilibrium binding constant [103]. Nonetheless, dynamic off-rates of desA-NDSK ($8.6 \times 10^{-4}s^{-1}$) and desAB-NDSK ($1.35 \times 10^{-3}s^{-1}$) from fibrinogen have been calculated from bond-strength measurements recorded with laser tweezers-based force spectroscopy [81]. Surprisingly, these off-rates are similar to the k_ds measured for the knob peptides in the present study ($\sim 10^{-3}s^{-1}$). Although there are inherent differences between the experimental parameters of SPR and laser tweezers-based force spectroscopy, the agreement of dynamic rates may provide further evidence that the knob:hole interaction is predominately mediated by the first few residues on the knob N-termini.

Performing MD simulations of the peptides in an aqueous environment lent substantial insight into potential peptide structural conformations and intrachain interactions that may influence binding properties. We acknowledge, however, the basic limitations of MD simulations: conformational sampling and the energy function. Moreover, we used theoretical models as starting structures, based on the published active conformations of GPRP and GPRV, because the structures for the peptide variants used in this study cannot be determined experimentally. Additional modeling would need to be performed to fully address these concerns. However, on the basis of our simulations, we noticed 2 main characteristics that correlated with functional binding affinities, first, the orientation of the 3Arg side chain, and second, backbone stability. For the superior binding peptide GPRPFPAC, the 3Arg side chain χ_1 angle was maintained in the gauche$^+$ rotameric conformation, and the weighted average of the backbone RMSD from the active conformation was the lowest. Meanwhile, the 3Arg group in the -ERC peptides (GPRPPERC and GPRVVERC) was stabilized in the gauche$^+$ conformation by a salt bridge ionic interaction between 3Arg and 6Glu, thereby also stabilizing the backbone. However, the salt bridge may initially restrict 3Arg from interacting with residues on the hole domains. Surprisingly, the 3Arg side chain

for GPRPAAC, a high-affinity peptide, was predominately observed in the trans conformation; a similar 3Arg rotameric conformation was observed for GPRVVAAC. Notably, the peptides used in the experimental and subsequent MD simulations were not amidated so as to facilitate future conjugation chemistries. In doing so, a negative charge was generated at the carboxyl-terminus, allowing presumably weak ionic interactions between a nonrestricted positively charged Arg side chain and the C-terminus as observed with GPRPAAC and GPRVVAAC. Despite this, the RMSD of the backbone for GPRPAAC was less than GPRVVAAC, indicating that the 4Pro residue probably stabilizes the peptide backbone. The resulting electrostatic potential was then subsequently influenced by the structural conformations of the side chains. This modeling analysis ultimately suggested that the conformation of knob peptides within an aqueous environment before contact with a hole domain contributes to the propensity of binding that occurs. This theory may translate to the N-terminal knobs on the more complex native fibrin monomer; however, significant additional experimentation would be required.

Using molecular dynamic simulations coupled with experimental SPR, we report for the first time, to our knowledge, fundamental knob structural determinants that drive knob:hole binding dynamics and provide potential knob design criteria. Exemplifying these criteria, we report a novel knob mimic with enhanced affinity. Collectively, we believe our studies provide additional insights for developing higher affinity peptides that bind and disrupt the native knob:hole interaction more effectively than previously reported knob peptides.

2.6 Acknowledgments

We thank Dr Shuming Nie for access to and assistance with the Biacore 2000. We also thank Dr Steven Harvey for the intellectual discussions and insights on molecular modeling and dynamics. The MD work was performed with the use of the computational resources of the Interactive High Performance Computing Laboratory at Georgia Institute of Technology.

This work was supported by the W. H. Coulter Foundation (GTF125000120) and the National Institutes of Health (1R21EB008463; T.H.B.) and by the NIH FIRST Post-Doctoral Fellowship (K12 GM000680; S.E.S.).

Table 2: Langmuir 1:1 model, fitted parameters.

Parameter	GPRPAAC	GPRPFPAC	GPRPPERC	GPRVVERC	GPRVVAAC
B_{max} (RU)	21.47 ± 0.19 *	17.74 ± 0.08	19.20 ± 0.21	16.27 ± 0.08	18.88 ± 0.18
k_a $(M^{-1}s^{-1}) \times 10^{-3}$	1.14 ± 0.03	12.5 ± 0.26	1.29 ± 0.03	0.28 ± 0.004	0.09 ± 0.001
k_d $(s^{-1}) \times 10^3$	11.71 ± 0.17	7.27 ± 0.08	9.92 ± 0.13	23.43 ± 0.23	10.73 ± 0.04
RSS	2.197	2.092	1.885	1.521	1.497

B_{max} indicates maximal binding capacity of D fragment; RU, resonance unit; k_a, association rate; k_d, dissociation rate; RSS, residual sum of squares.

* Mean ± SD (all such values).

Table 3: Heterogeneous ligand model, fitted parameters.

Parameter	GPRPAAC	GPRPFPAC	GPRPPERC	GPRVVERC	GPRVVAAC
B_{max} (RU)	$19.38 \pm 0.23^{*}$	17.59 ± 0.13	14.58 ± 0.23	25.61 ± 0.28	8.36 ± 0.24
k_{a1} $(M^{-1}s^{-1}) \times 10^{-3}$	2.84 ± 0.14	21.72 ± 0.73	3.22 ± 0.12	1.07 ± 0.03	0.62 ± 0.04
k_{d1} $(s^{-1}) \times 10^{-3}$	12.83 ± 0.49	81.10 ± 2.23	58.73 ± 1.86	57.67 ± 2.24	30.08 ± 1.51
k_{D1} $(\mu M)^{\dagger}$	4.53	3.73	18.23	53.89	48.51
B^{*}_{max} (RU)	5.91 ± 0.08	11.34 ± 0.08	5.92 ± 0.13	10.33 ± 0.23	13.52 ± 0.32
k_{a2} $(M^{-1}s^{-1}) \times 10^{-3}$	1.05 ± 0.09	1.81 ± 0.04	1.01 ± 0.03	0.26 ± 0.01	0.04 ± 0.001
k_{d2} $(s^{-1}) \times 10^{-3}$	8.95 ± 1.05	1.96 ± 0.04	2.96 ± 0.09	4.07 ± 0.06	1.00 ± 0.07
k_{D2} $(\mu M)^{\dagger}$	8.52	1.08	2.93	15.71	25.00
RSS	1.474	0.9887	0.9437	0.9817	1.266

B_{max} indicates maximal binding capacity of D fragment; RU, resonance unit; k_a, association rate; k_d, dissociation rate; RSS, residual sum of squares.

* Mean $\pm$ SD (all such values).

† Calculated K_D from fitted k_a and k_d values in which $K_D = k_d/k_a$.

CHAPTER III

COMPUTATIONAL SCREENING AND DESIGN OF DNA-LINKED MOLECULAR NANOWIRES[1]

3.1 Abstract

DNA can be used as a structural component in the process of making conductive polymers called nanowires. Accurate molecular models could lead to a better understanding of how to prepare these types of materials. Here we present a computational tool that allows potential DNA-linked polymer designs to be screened and evaluated. The approach involves an iterative procedure that adjusts the positions of DNA-linked monomers in order to obtain reasonable molecular geometry compatible with normal DNA conformations and with the properties of the polymer being formed. This procedure has been used to evaluate designs already reported experimentally, as well as to suggest a new design based on pyrrylene vinylene (PV) monomers.

3.2 Main

DNA is playing an ever-increasing role in creating new nanoscale materials and devices with applications in areas such as electronics and biomedicine. Seeman and co-workers, for instance, developed a way to form short segments of a nylon-like polymer covalently linked to a nucleic acid backbone [170] with the long-term goal of "controlling the topology of industrial polymers by nucleic acids" [83]. More recently, Datta et al. reported a novel DNA-mediated technique for forming short oligomers of the conducting polymer polyaniline (PANI) [37]. Exploiting the sequence programmability of DNA, they prepared 22-mer oligonucleotides designed with a contiguous stretch of six modified cytosine bases in the

[1]Reprinted with permission from Gossett, J. J. and Harvey, S. C., "Computational screening and design of DNA-linked molecular nanowires," *Nano Lett.*, vol. 11, no. 2, pp. 604–608, 2011. Copyright © 2010 American Chemical Society.

middle of the sequence. Each of these modified bases contained an aniline moiety covalently linked at the N4 nitrogen with a $-(CH_2)_2-$ linker. After the modified oligonucleotides were hybridized with complementary strands to form duplexes, aniline polymerization was achieved by oxidizing the duplexes with H_2O_2 and horseradish peroxidase. Despite evidence of significant structural distortion, these DNA-linked oligoaniline structures were found to have the properties of a conducting polymer, a result seen as encouraging for the development of nanowires [37].

The method introduced by Datta et al. was later expanded to prepare oligomers of poly(4-aminobiphenyl) (PAB) covalently linked to DNA [36]. In contrast to the aniline monomers, however, the 4-aminobiphenyl monomers were placed at every other nucleobase in the sequence, because 4-aminobiphenyl contains one more aromatic ring than aniline. Further demonstrating the range of polymers that can be formed by this technique, Srinivasan and Schuster synthesized poly-thienopyrrole (TP) oligomers with the thienopyrrole monomers placed on every other nucleobase, this time using a $-(CH_2)_3-$ linker instead of a $-(CH_2)_2-$ linker [135].

This approach is not, however, without limitations. Chen et al. pointed out that the three-dimensional structure of PANI is "incommensurate" with that of DNA because PANI does not adopt a helical conformation with the same rise and twist as typical B-form DNA [29]. (Rise is the translation along the helix axis per repeat unit and twist is the rotation about the axis per repeat unit.) They argue that this creates distortions in the DNA scaffold, limiting the potential length of the PANI polymer. In contrast, their experiments with 2,5-bis(2-thienyl)pyrrole (SNS) monomers indicated that the resulting SNS oligomers were "structurally commensurate with duplex DNA" and therefore could in principle be made of any length.

Whether or not a polymer can adopt a helical conformation that is structurally commensurate with DNA will depend on the bond lengths, bond angles, and dihedral angles of the repeating unit, i.e., the monomer. Thus, determining that the polymer can adopt a helical conformation with the same rise and twist as DNA is necessary *but not sufficient* for evaluating whether the polymer will be commensurate with DNA. One must also consider the

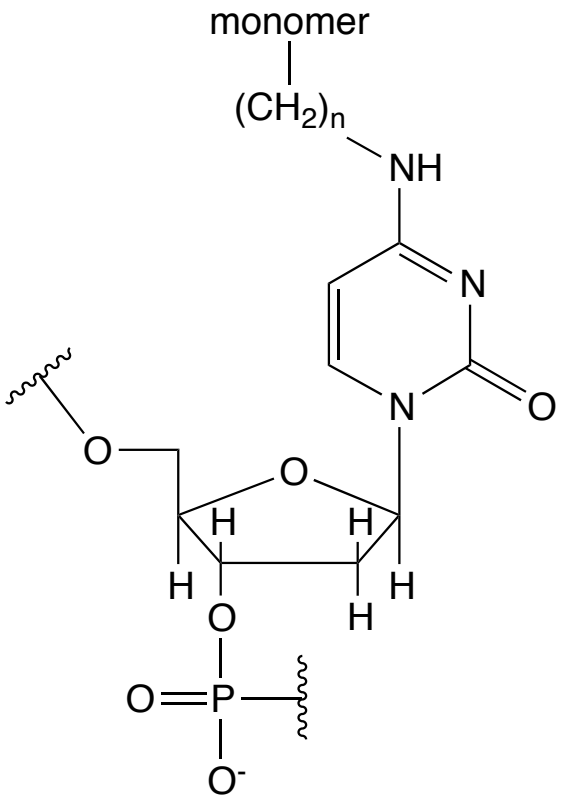


Figure 6: Cytosine base modification for preparing DNA-linked polymers. Monomers are attached at the N4 nitrogen by a short chain of C–C single bonds.

steric compatibility of the polymer and the DNA when choosing the linker, the monomer attachment point, and the attachment interval on DNA (i.e., every base or every other base). Molecular modeling can help the experimentalist make these decisions in order to minimize structural distortion of the DNA duplex after the monomers are coupled together. In the present study, our goal was to develop a modeling strategy that explicitly takes into account how the monomer is connected to DNA—a strategy that allows one to screen and evaluate potential designs for DNA-assisted polymer synthesis, potentially saving time and experimental effort.

Our approach was to build a model with monomers covalently linked to DNA and then to evaluate whether the monomers could be coupled together to form a polymer without overly distorting the DNA structure. Essentially this is a loop closure problem. A "loop" in this context is defined as two successive monomers each connected to DNA that are linked together by a covalent bond. To adjust the positions of the monomers so that standard

bond lengths and angles might be obtained for the covalent bonds between monomers, we implemented a loop closure method based on the cyclic coordinate descent (CCD) algorithm [146]. Canutescu and Dunbrack have previously applied this algorithm to the protein loop closure problem [26]. Briefly, they built a protein loop by anchoring the N-terminal end of the loop to the protein model, allowing the other end to move freely. The CCD algorithm involved cycling through each backbone dihedral angle of the loop in order to connect the moving end to the fixed target residue of the protein model. Only one degree of freedom was adjusted at a time, and they were able to derive an analytical expression for the optimal change for the dihedral angle. In general, many cycles were needed to achieve loop closure. Our loop closure problem presented some unique challenges, detailed below.

Starting with a fixed all-atom representation of double-helical DNA, two monomers are added to form a loop. Any DNA conformation can be used as long it has helical symmetry, but the conformation remains fixed throughout the procedure. Depending on the design, the monomers are attached to successive nucleobases or by skipping one or more bases between modifications. In the present application, each monomer is linked to DNA at the N4 nitrogen of a cytosine base with a short chain of C–C single bonds known as the linker (Figure 6). The linker length, the attachment interval, and the attachment position are parameters of the design. Three dummy atoms are then added to the first monomer, and closure is monitored by measuring the overlap between these and three corresponding real atoms of the second monomer while holding the DNA conformation fixed (Figure 7). The overlap is defined by the root-mean-square deviation (rmsd), given by

$$\text{rmsd} = \sqrt{\frac{1}{3}\sum_{i=1}^{3}\delta_i^2} \tag{2}$$

where δ_i is the distance between the ith dummy atom and the ith real atom.

We use an internal coordinate representation for the linker and monomer, i.e., a bond length, a bond angle, and a dihedral angle locate each atom. Bond lengths and bond angles remain fixed throughout the calculations. The dihedral angles that determine the initial conformation are assigned random values. Equating symmetrically equivalent dihedral angles imposes helical symmetry. Conversion from internal coordinates to Cartesian (x, y, z)

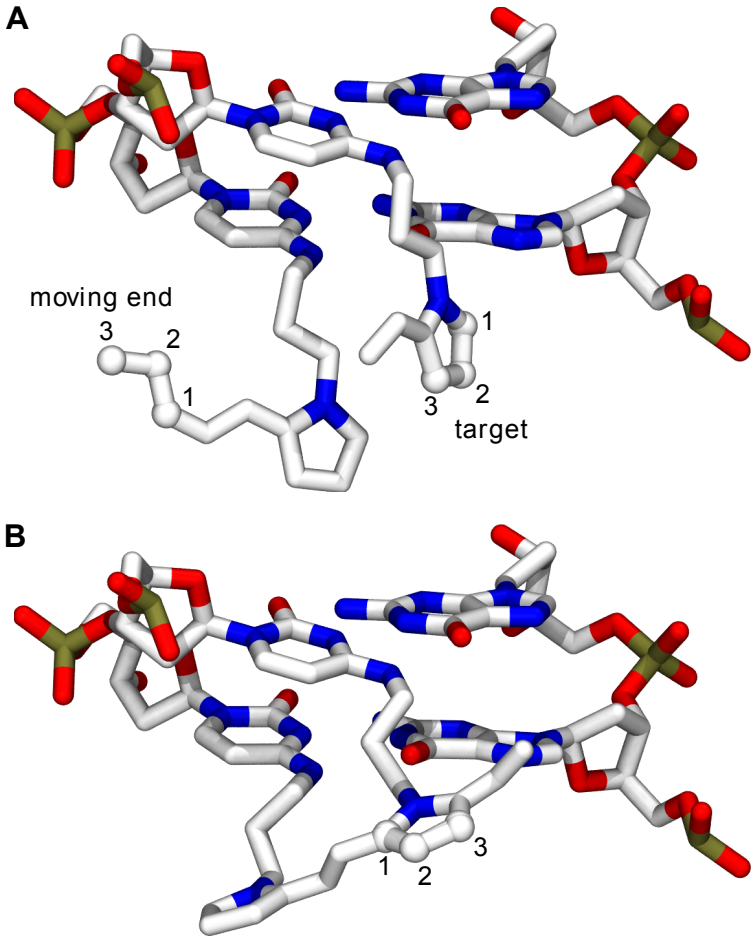


Figure 7: Pyrrylene vinylene (PV) monomers attached to two consecutive cytosine bases, forming a loop. The loop closure algorithm attempts to minimize the overlap between the three dummy atoms attached to the first monomer (moving end) and the three real atoms of the second monomer (target). (A) Conformation prior to running the loop closure algorithm. (B) Conformation after the algorithm has stopped.

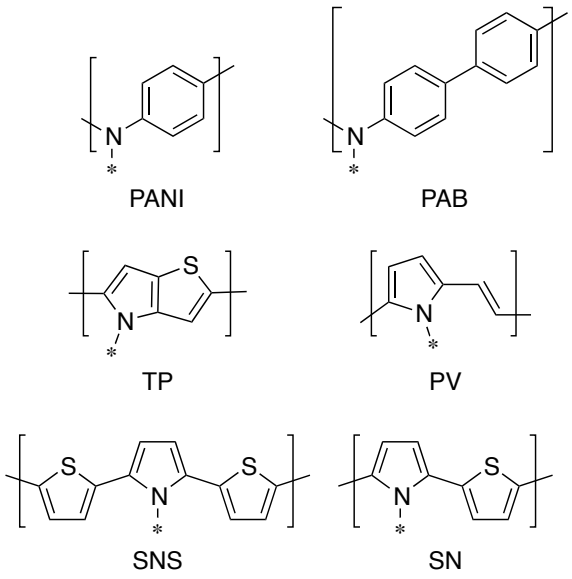


Figure 8: Polymers evaluated in this study. Attachment points are indicated. Abbreviations are: PANI, polyaniline; PAB, poly(4-aminobiphenyl); TP, poly(thienopyrrole); PV, poly(pyrrylene vinylene); SNS, poly(2,5-bis(2-thienyl)pyrrole); SN, poly(thiophene-pyrrole).

coordinates is performed using the SN-NeRF algorithm [112].

Much like the CCD algorithm reported by Canutescu and Dunbrack, an iteration of our implementation involves cycling through the dihedral angles of the model, adjusting one angle at a time, to minimize the closure rmsd [26]. To determine the optimal amount to change each dihedral angle, however, we use a numerical method. This is because the target for our loop is not fixed in space like it is for protein loops. In our system the three atoms of the second monomer on which we are trying to superimpose the three dummy atoms of the first monomer define the target. Since the dihedral angles of the second monomer are set equal to the dihedral angles of the first monomer in order to impose helical symmetry, the target is moving simultaneously. Thus, we use a one-dimensional minimization algorithm called Brent's method [20] to calculate the optimal value for each dihedral angle. The loop closure algorithm will stop when the rmsd falls below 0.08 Å, indicating successful loop closure, or when the number of CCD iterations exceeds a prespecified limit (typically 5000) [26]. We make no attempt to avoid steric clashes while running the algorithm; clashes can be checked after the algorithm has stopped.

We evaluated each of the polymers shown in Figure 8. To reduce the bias in our results

we used three different DNA structures, and, since our method is susceptible to local optima and steric clashes, we ran the algorithm 100 times for each of the DNA conformations. The selection of DNA structures is important because the DNA remains fixed throughout the procedure. To reduce the risk of an incorrect result, one should choose a subset of DNA structures that are in some sense representative of DNA conformational space. The DNA structures (A-, B-, and C-form DNA) were obtained using the Web 3DNA web server [169]. We used a poly(dG)–poly(dC) sequence with ideal geometry for each DNA structure. A-DNA was constructed with rise = 2.548 Å and twist = 32.7°, B-DNA was constructed with rise = 3.375 Å and twist = 36.0°, and C-DNA was constructed with rise = 3.310 Å and twist = 38.6° [169]. Attachment designs are defined in Table 4.

Table 4: Minimum RMSD Obtained for 100 Loop Closure Trials.

				minimum RMSD (Å)		
	polymer	linker length[a]	attachment interval[b]	A-DNA	B-DNA	C-DNA
1	PANI	2	1	2.600	1.434	1.286
2	PAB	2	2	3.474	1.201	0.978
3a	TP	3	1	1.007	0.152	0.190
3b	TP	3	2	0.182	0.863	0.948
4a	SNS	2	2	2.159	*<0.08*	*<0.08*
4b	SNS	3	2	1.193	*<0.08*	*<0.08*
5	PV	3	1	1.089	*<0.08*	0.090
6a	SN	3	1	2.420	0.822	0.478
6b	SN	4	1	1.872	0.294	0.086
6c	SN	5	1	1.222	*<0.08*	*<0.08*

[a] Number of carbon atoms in the linker.
[b] Key: 1, every base; 2, every other base.

Loop closure was achieved with DNA-linked polymer designs **4a**, **4b**, **5**, and **6c** (Table 4). For design **4b** (SNS) with B-form DNA, loop closure was successfully achieved in all 100 trials. Loop closure was successfully achieved in 33 out of 100 trials for design **5**

(pyrrylene vinylene, or PV) with B-form DNA. Just because loop closure was achieved, however, does not mean that the structure is plausible. The structure must also be checked for unreasonable steric clashes. Using a nonbonded rejection distance of 1.7 Å, we found that out of the 100 successful loop closures achieved using design **4a**, seven were acceptable, and out of 33 successful loop closures achieved using design **5**, four were acceptable. Steric clashes with contact distances on the order of 1.7 Å can be repaired by energy minimization or annealing of the entire system.

Polymers of SNS and PV monomers, respectively, can form a helical conformation with the same rise and twist as B-DNA, partially explaining the results we obtained. This can be shown by calculating the rise and twist of the polymer (independent of DNA) for different values of the torsion angles in the repeating unit [137]. Briefly, we used a systematic (or grid) search to alter the rotatable torsions to provide adequate coverage of conformational space, and then we determined whether the rise and twist of a given DNA conformation coincided with the region of possible rise and twist values calculated for the polymer (data not shown). As mentioned above, however, the attachment configuration is also critical, as demonstrated by the thiophene-pyrrole (SN) polymer. It can adopt a helical conformation that has the same rise and twist as, say, B-form DNA, but loop closure was not achieved for 3- and 4-carbon linkers. This is because the helical conformation results in a structure where the polymer is displaced far from the helix axis, and therefore a longer linker would be needed.

The results for PANI, PAB, and TP are not surprising, because these polymers do not form a helical conformation with the same rise and twist per repeat unit as any of the DNA conformations that were used for testing. Loop closure was simply not possible. Another way of looking at this is that the repeating units for the PANI, PAB, and TP polymers have only one independent torsional degree of freedom. (PV and SNS, on the other hand, have two and three independent torsion angles, respectively.) With only one independent degree of freedom, one cannot generally satisfy two restraints; in this case one cannot generally match the rise and helical twist of DNA. More torsional degrees of freedom make it more likely that the polymer can adopt a helical conformation that is compatible with the DNA

helix. Nonetheless, running the loop closure test in these cases is still worthwhile as can be seen by the following example. Despite the failure of TP polymers to form a compatible helix, it is interesting to note that loop closure was nearly achieved using A-DNA with design **3b**. Srinivasan and Schuster had some success forming TP oligomers from DNA containing six TP units attached in this configuration; however, melting temperature experiments suggested that the duplex structure was distorted [135]. Our loop closure results suggest that TP monomers connected using design **3a** might be better for maintaining a B-form duplex than using design **3b**, even though loop closure was not achieved in either case.

Our loop closure screening method depends on several assumptions, including helical symmetry and fixed DNA. To determine if our assumptions were reasonable and to study the stability of the DNA-linked polymers, we performed classical molecular dynamics (MD) simulations on DNA-linked polymers with a subset of the designs from Table 4 (specifically, designs **1**, **2**, **3a**, **3b**, **4b**, and **5**). We used 19-bp DNA duplexes to ensure the polymer would extend for at least one full turn. Monomers were added to bases as specified by the attachment configuration, but monomers were not attached to the first and last bases of the modified strand. The monomers were attached using a conformation that came from a successful loop closure trial, or a manually chosen conformation if loop closure was unsuccessful. Coordinate and parameter-topology files were prepared using the programs *antechamber* and *LEaP* [27]. Each system was solvated with TIP3P water molecules, and then we added 36 K^+ ions to achieve electric neutrality. Energy minimization and molecular dynamics simulations were performed with NAMD Version 2.7b3 [115] using AMBER force field parameters [145]. Bonds to hydrogen atoms were constrained using the SHAKE algorithm [121]. Short-range nonbonded interactions were cut off with a switching function between 10 and 12 Å, while long-range electrostatic interactions were calculated using the particle mesh Ewald (PME) method [35]. First, we ran 1000 steps of energy minimization with the solute atoms fixed and then another 1000 steps of minimization with no atoms fixed. Next, with harmonic restraints on the solute atoms, the systems were heated from 0 to 310 K over a time period of 20 ps, and then equilibrated at 310 K and 1 atm pressure for 100 ps. After the harmonic restraints were removed, the systems were equilibrated for an

additional 200 ps. Finally, we ran 10 ns simulations under constant temperature (310 K) and pressure (1 atm) conditions. We used a 2 fs integration time step. Atomic coordinates were saved every 1 ps for analysis.

DNA stability was assessed by the time evolution of the root-mean-square deviation (rmsd) of the nucleic acid atoms (Figure 9A). The reference structure for calculating the rmsd for each system was the first structure after the harmonic restraints were removed. Not surprisingly, the rmsd of the systems with designs **1** and **2** was much higher than the rmsd of the systems with designs **4b** and **5**. Snapshots of the DNA–polymer complexes at the end of the 10 ns simulations are shown in Figure 9B. The structures based on designs **1** and **2** are severely distorted, while the duplexes based on designs **4b** and **5** maintain a B-form-like conformation. This result suggests that the PV polymer, like SNS, is structurally commensurate with DNA. The structure with design **3b** is interesting in that the TP oligomer runs straight up the helix axis, yet the base pairing remains mostly intact. The rmsd between this structure and ideal B-form DNA is 8.6 Å, whereas the rmsd between this structure and ideal A-form DNA is 6.8 Å, which would suggest more similarity with A-form than B-form, but perhaps not significantly so. Putting the TP monomers on every base (design **3a**) results in a much different structure, a structure resembling an elongated B-form duplex. Both compounds **3a** and **3b** resulted in plausible DNA structures, demonstrating that just because loop closure is not achieved does not necessarily mean that a design should be rejected outright.

Importantly, these simulation results are consistent with our loop closure results. Those designs where loop closure was achieved were more likely to maintain typical DNA conformations during the simulations. Furthermore, the designs resulting in a higher minimum closure rmsd corresponded to significantly greater distortion during the simulations. In the case of design **3a**, the loop closure tests hinted that the structure might resemble a B-form conformation, and indeed the structure predicted from the simulation suggested an elongated B-form duplex. Similarly, the loop closure tests hinted that **3b** might resemble an A-form conformation, and indeed the structure predicted from the simulation suggested a plausible A-form structure.

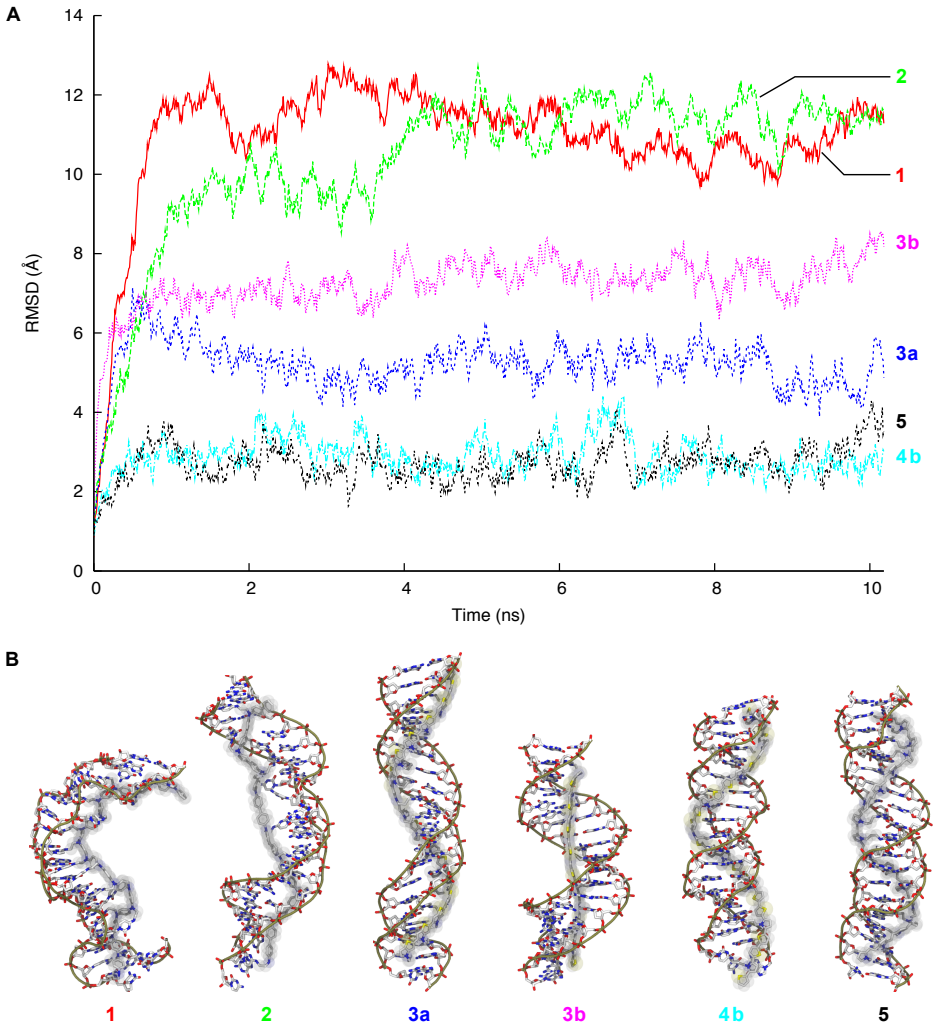


Figure 9: Results of MD simulations. (A) Root mean square deviation (rmsd) of the nucleic acid atoms as a function of simulation time. (B) Snapshots of the DNA–polymer complexes at the end of the simulations with the polymers highlighted. Snapshots rendered using VMD Version 1.8.7 [62].

In summary, our method provides a convenient and systematic way to screen potential monomers and attachment designs. The assumption underlying our approach is that efficacy of loop closure is indicative of the likelihood of a polymer being formed with minimal distortion of the DNA structure when DNA modified with the monomers is oxidized to form the DNA–polymer complex. If it is indeed the case that DNA distortion limits the potential length of the polymer, as has been previously suggested [29], then our method could help the experimental community create new designs for DNA-linked polymer nanowires. In this study, we have identified PV as a good candidate for DNA-assisted polymer synthesis, as well as found an alternative attachment configuration for TP. As experimentalists expand the technique to include different attachment points on DNA, or to create heteropolymers, this screening and design method will prove to be a valuable tool.

3.3 Acknowledgments

We thank Gary Schuster for suggesting this problem. Loop closure trials were performed using the computational resources of the Interactive High Performance Computing Laboratory (IHPCL) at the Georgia Institute of Technology. Supported by the Georgia Research Alliance.

CHAPTER IV

DOMAIN III OF THE *T. THERMOPHILUS* 23S rRNA FOLDS INDEPENDENTLY TO A NEAR-NATIVE STATE[1]

Abstract: The three-dimensional structure of the ribosomal large subunit (LSU) reveals a single morphological element, although the 23S rRNA is contained in six secondary structure domains. Based upon maps of inter- and intra-domain interactions and proposed evolutionary pathways of development, we hypothesize that Domain III is a truly independent structural domain of the LSU. Domain III is primarily stabilized by intra-domain interactions, negligibly perturbed by inter-domain interactions, and is not penetrated by ribosomal proteins or other rRNA. We have probed the structure of Domain III rRNA alone and when contained within the intact 23S rRNA using SHAPE (selective 2'-hydroxyl acylation analyzed by primer extension), in the absence and presence of magnesium. The combined results support the hypothesis that Domain III alone folds to a near-native state with secondary structure, intra-domain tertiary interactions, and inter-domain interactions that are independent of whether or not it is embedded in the intact 23S rRNA or within the LSU. The data presented support previous suggestions that Domain III was added relatively late in ribosomal evolution.

4.1 Introduction

The ribosome is our most direct macromolecular connection to the distant evolutionary past and to early life [160, 162, 132, 17, 60, 11, 48]. The ribosome is believed to have emerged from the "RNA world" [119, 159, 34, 109, 51] following an evolutionary pathway that preserved ribosomal RNAs as central players in peptide bond formation and decoding

[1]This chapter was adapted from Athavale, S. S., Gossett, J. J., Hsiao, C., Bowman, J. C., O'Neill, E., Hershkovitz, E., Preeprem, T., Hud, N. V., Wartell, R. M., Harvey, S. C., and Williams, L. D., "Domain III of the T. thermophilus 23S rRNA folds independently to a near-native state," *RNA*, vol. 18, no. 4, pp. 752–758, 2012, doi: 10.1261/rna.030692.111 (`www.rnajournal.org`).

[106, 10, 104, 56, 108, 167, 129, 130]. Understanding the origin and evolution of rRNA is a key to understanding the early evolution of life on earth.

The ribosome is made of a small subunit (SSU) and a large subunit (LSU). The SSU in bacteria and archaea contains a single RNA molecule, the 16S rRNA. Phylogenetic studies by Woese et al. [161] revealed three major and one minor secondary structural domains (2° domains) of the 16S rRNA. These 2° domains are segregated into independent and autonomous three-dimensional domains (3D domains) in the assembled SSU. Each 2° domain of the 16S rRNA folds and assembles with the appropriate ribosomal proteins into a 3D domain, independent of other 2° domains [152, 123, 1]. One 3D domain is called the head and others are called the body and the platform [21, 157]. The head, body, and platform domains of the SSU have direct functional significance, moving independently during translation [105]. These 3D domains may also have evolutionary significance. The domain is the evolutionary unit of protein evolution [24, 47]. Protein domains are modular units that are combined and recombined over evolution to achieve various functions. It is conceivable that the 3D domains of the SSU played analogous evolutionary roles, but on a more ancient timeframe. If so, then the 3D domains of the SSU may have been recruited to the ribosome, from prior functional roles.

The LSU in bacteria and archaea is made up of a 23S rRNA and a much smaller 5S rRNA. The 23S rRNA contains six 2° domains (Fig. 10A; [107]). Although these 2°domains are well-defined in the secondary structure, in three dimensions the LSU appears monolithic [139, 10, 167]. It has been suggested that, unlike in the SSU, the 2° domains in the LSU do not correspond to 3D domains.

Questions naturally arise as to whether the architectures and early evolution of the SSU and the LSU are fundamentally different, and if so, why? Do isolated 2° domains of the 16S rRNA but not the 23S rRNA act as 3D domains and fold to near-native 3D structures? How are the 2° domains of the 16S and 23S rRNAs related to 3D structure, function, and evolution of the ribosome?

Here we experimentally probe the domain structure of the LSU. We show that one isolated 2° domain of the 23S rRNA can fold to a near-native state in absence of the

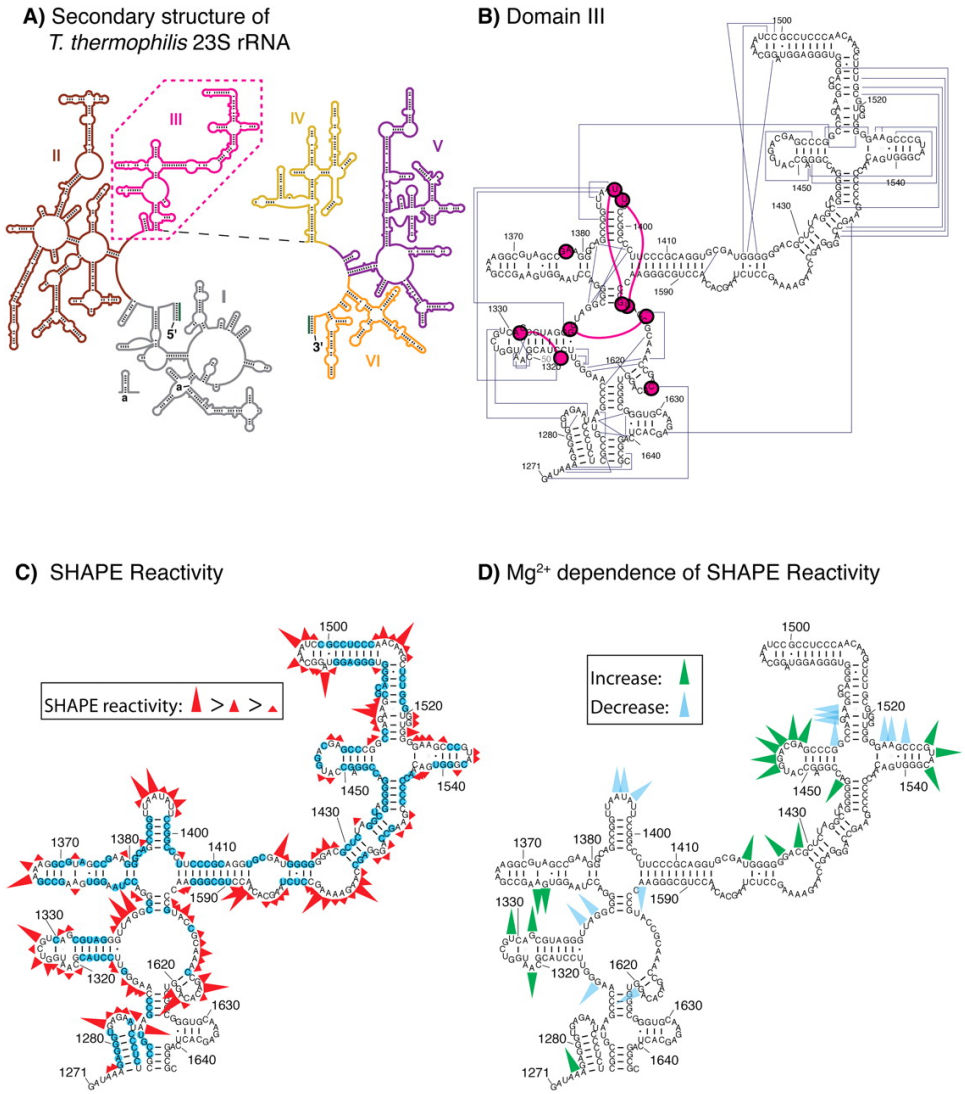


Figure 10: (A) Secondary structure of the 23S rRNA of the large subunit of *T. thermophilus* (adapted with permission from Harry Noller). The six secondary structural domains of 23S rRNA are shown: Domain I in gray, Domain II in brown, Domain III in pink, Domain IV in yellow, Domain V in purple, and Domain VI in orange. (B) Tertiary interactions (dark blue) and phosphate–magnesium–phosphate linkages within Domain III. Each first shell magnesium–phosphate interaction is indicated by a magenta circle. The lines between the circles are the phosphate–magnesium–phosphate linkages. (C) SHAPE reactivities for Domain III$^{\text{alone}}$ in 250 mM Na^{+}. The blue nucleotides are unreactive. (D) Magnesium-dependent SHAPE reactivities for Domain III$^{\text{alone}}$, observed upon addition of 10 mM Mg^{2+}. Only the nucleotides with the greatest proportional change in reactivity are indicated.

remainder of the LSU, and appears to be a true 3D domain. Our focus here is Domain III of the *Thermus thermophilus* 23S rRNA (Fig. 10B), which is described by Thirumalai and colleagues [64] as compact and slightly prolate. We use SHAPE [93, 154] to demonstrate that Domain III excised from the 23S rRNA (Domain III$^{\text{alone}}$) folds in a magnesium-dependent fashion to the same basic state as when it is embedded in the intact 23S rRNA (Domain III$^{\text{23S}}$). In this near-native state of Domain III, surface residues appear to be poised with the correct geometry for the inter-domain rRNA–rRNA interactions observed in the structure of the LSU (PDB entry 2J01) [130]. Our results are consistent with the structure of Domain III within the LSU where Domain III is compact, and its interactions with other ribosomal components are restricted to its surface (Figs. 11, 12; Fig. 29 in Appendix B).

4.2 Results

4.2.1 SHAPE accurately predicts the canonical secondary structure of Domain III$^{\text{alone}}$

The canonical secondary structure of the 23S rRNA, based on comparative sequence analysis [167, 25], is strongly supported by previous SHAPE experiments [38]. As shown by Weeks and colleagues, SHAPE exploits the reactivity of the 2'-hydroxyl groups of RNA with electrophilic chemical probing reagents such as NMIA (N-methylisatoic anhydride) or BzCN (benzoyl cyanide) [93, 154]. The relative reactivities of the 2'-hydroxyl groups of various nucleotides are sensitive primarily to local RNA flexibility. Consequently, paired nucleotides within helical regions are generally less flexible and less reactive toward SHAPE reagents than unpaired nucleotides.

The close correspondence of our SHAPE data to the canonical secondary structure of Domain III is illustrated in Figure 10C, where SHAPE reactivity of Domain III$^{\text{alone}}$ is mapped onto the canonical secondary structure. All SHAPE reactivity data were obtained using NMIA unless otherwise specified. The definition of Domain III used here is conventional and includes residues G1271–G1647 of the 23S rRNA by the Escherichia coli numbering scheme [23, 167]. These data were obtained in presence of 250 mM Na^+ ions and in the absence of divalent cations. Under these conditions, RNA is expected to assume secondary structure but not necessarily tertiary structure [22, 42]. Consistent with this tendency, the

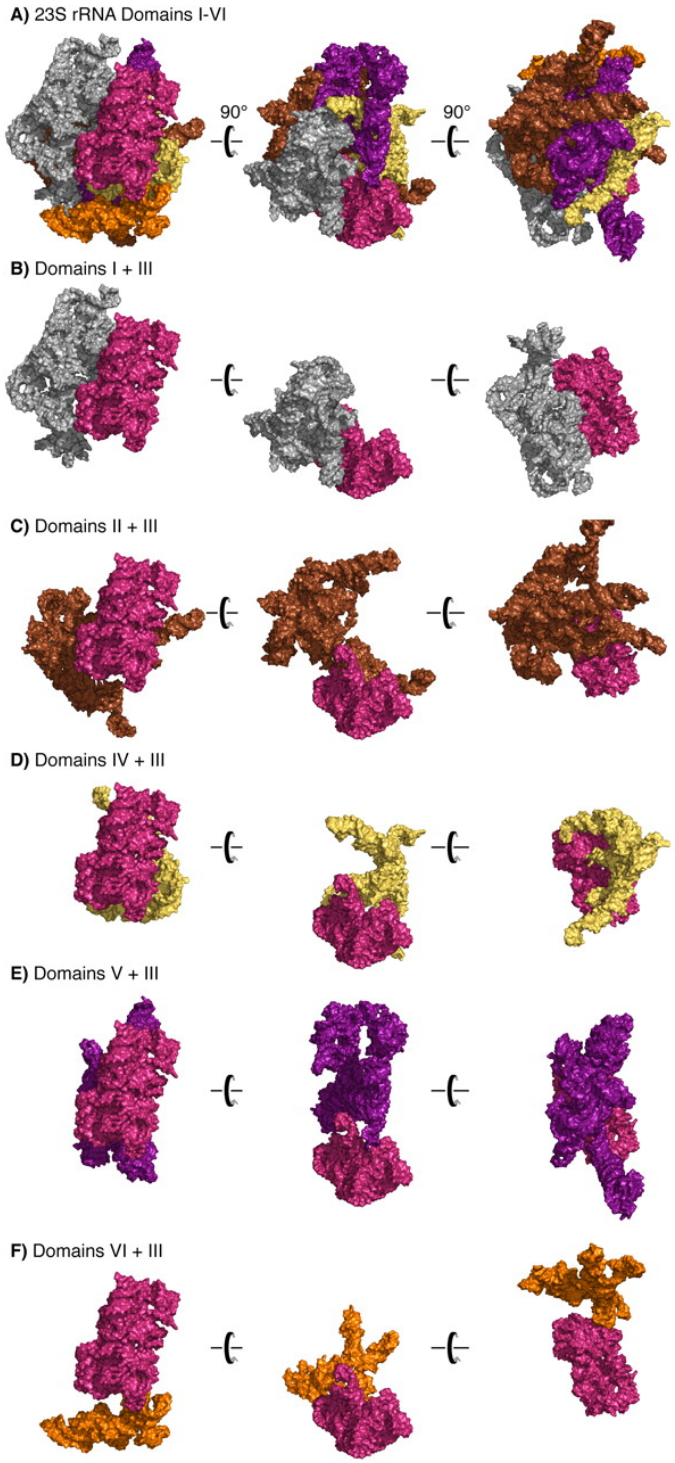


Figure 11: Domain III is compact and is not penetrated by other 2° domains. (A) All six 2° domains of the 23S rRNA are shown, colored as in Figure 10. Three views, with a relative rotation of 90°, are shown. (B–F) Interactions of Domain III with Domains I, II, IV, V, and VI, respectively.

correspondence between SHAPE reactivities and the secondary structure is very nearly perfect. Nucleotides of Domain III^{alone} were ranked using their absolute SHAPE reactivities relative to A1572 (highest reactivity) and binned into four groups, which are indicated in Figure 10C (see Appendix B for a more detailed analysis).

4.2.2 Folding of Domain III^{alone} to a near-native state requires magnesium ions

The folding of RNAs from secondary structure to their native states, containing long-range tertiary interactions, is known to be generally magnesium-dependent [22, 42]. The native state of Domain III rRNA, as inferred from the 3D structure of the assembled LSU, is stabilized by extensive networks of intra-domain tertiary base–base, base–backbone, and backbone–magnesium–backbone interactions (Fig. 10B). Consistent with this observation, Figure 10D shows that the magnesium-induced changes in SHAPE reactivity of Domain III^{alone} are widely dispersed over Domain III rRNA. The SHAPE reactivities increase at some sites and decrease at others. The nucleotides with SHAPE reactivities that are most sensitive to magnesium are mapped onto the secondary structure in Figure 10D. This magnesium dependence of the SHAPE reactivity reflects (i) specific magnesium binding, (ii) more diffuse interactions of magnesium with the RNA, and (iii) tertiary rRNA–rRNA intra-domain interactions (Tables 5, 6, Appendix B). Such magnesium-dependent SHAPE reactivity has previously been demonstrated for tRNA and RNase P [93, 99].

We used two chemical reagents to verify that the observed changes in reactivity are the result of RNA folding and not from direct modulation of the reagent activity by magnesium. Although SHAPE reactivity of NMIA has been shown to be modestly sensitive to magnesium [98], reactivity of BzCN is independent of magnesium [99]. We confirmed that NMIA and BzCN show similar changes in SHAPE reactivity upon addition of magnesium (Fig. 30, Appendix B).

4.2.3 The secondary structure of Domain III rRNA is conserved upon excision from the 23S rRNA

Figure 13A shows the SHAPE reactivities of Domain III^{alone} and Domain III^{23S}, both in the absence of magnesium. As illustrated by the overlaid traces, the reactivities are essentially

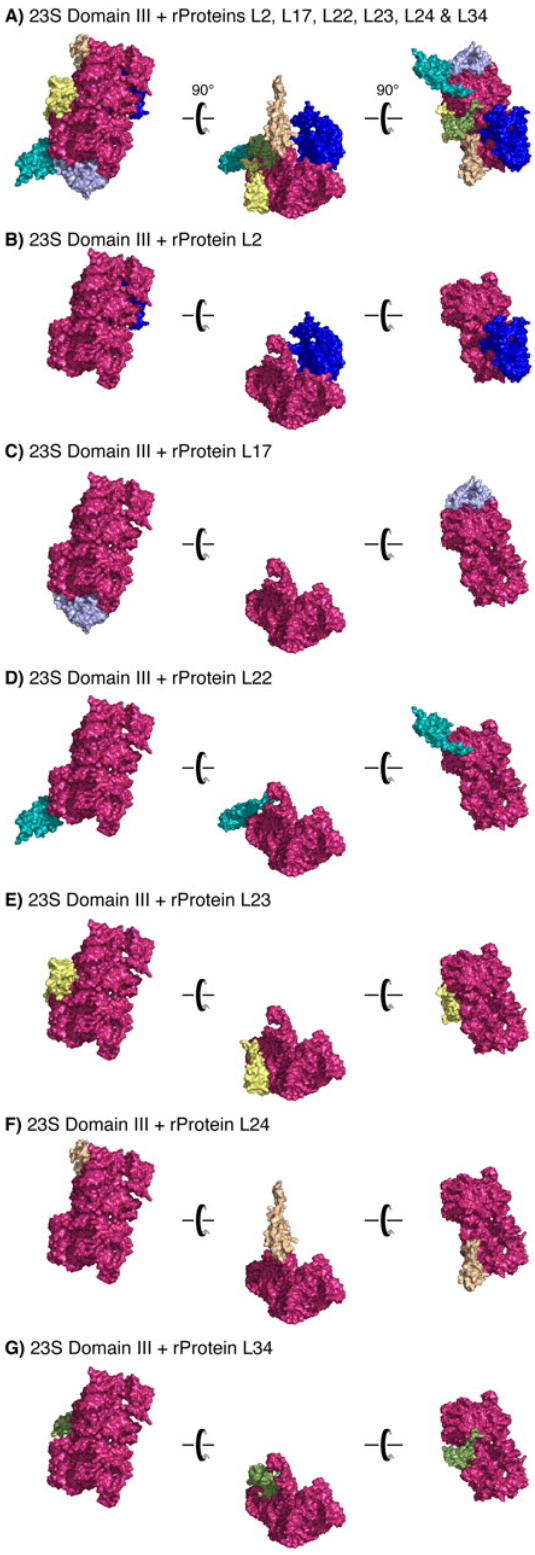


Figure 12: Domain III is not penetrated by ribosomal proteins. (A) Domain III, colored and oriented as in Figure 11, with rProteins L2 (dark blue), L17 (light blue), L22 (dark green), L23 (yellow), L24 (light brown), and L34 (light green). (B–G) Interactions of Domain III with each of these rProteins.

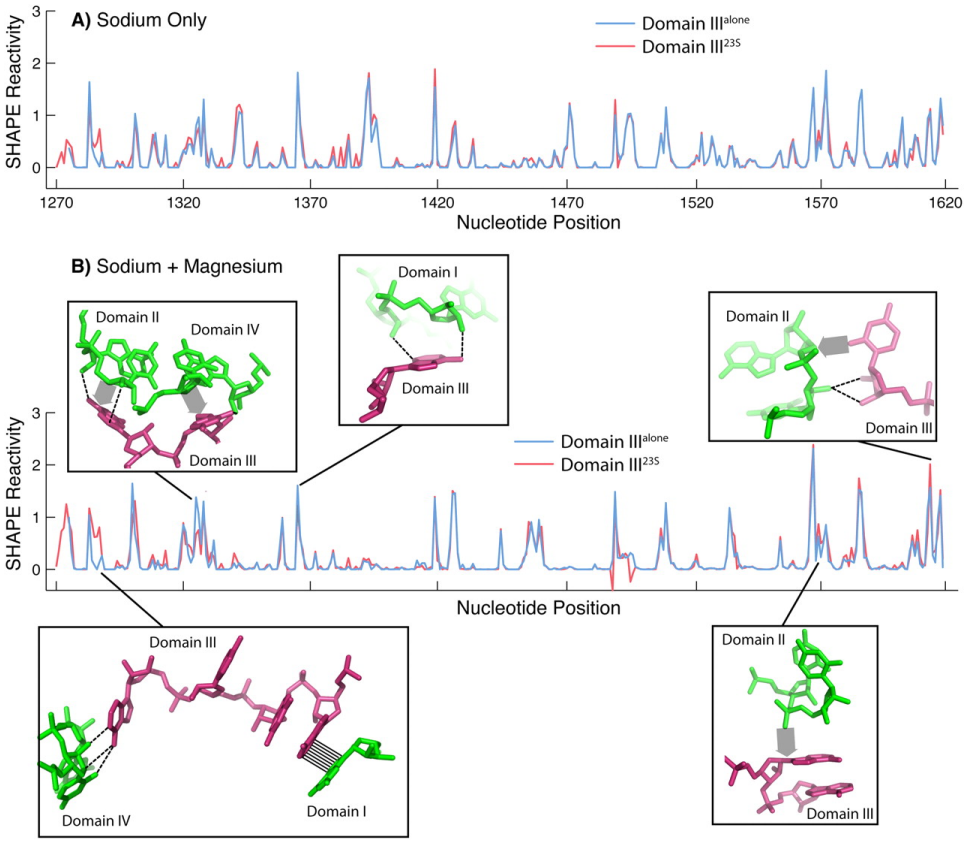


Figure 13: SHAPE reactivity for Domain III$^{\text{alone}}$ (blue) and Domain III^{23S} (red). The vertical axis represents SHAPE reactivities and the horizontal axis represents nucleotide position using conventional *E. coli* numbering scheme. (A) Domain III$^{\text{alone}}$ and Domain III^{23S} in 250 mM Na^+. (B) Domain III$^{\text{alone}}$ and Domain III^{23S} in 250 mM Na^+ and 10 mM Mg^{2+}. The inter-domain interactions between Domain III and Domains I, II, and IV that cause differences in SHAPE reactivity between Domain III$^{\text{alone}}$ and Domain III^{23S} are highlighted. Hydrogen bonds are shown by dashed lines, stacking interactions are shown by hashing, and van der Waals contacts are shown by broad shaded arrows.

identical along the length of the Domain III sequence. The high degree of similarity suggests that the secondary structure of Domain III^{alone} is the same as Domain III^{23S}.

4.2.4 Mg^{2+}-mediated folding of Domain III to the near-native state is conserved upon excision from the 23S rRNA

In the presence of magnesium ions, the SHAPE reactivities for Domain III^{alone} and Domain III^{23S} are very similar (Fig. 13B). The magnesium-dependent state of Domain III is therefore retained when it is excised from the 23S rRNA. The data also show that inter-domain rRNA–rRNA interactions are disrupted upon excision of Domain III from the 23S rRNA. In presence of magnesium, the SHAPE reactivity of Domain III^{23S} differs subtly from that of Domain III^{alone} (Fig. 13B). The differences are statistically focused at nucleotides involved in inter-domain interactions in the LSU, rather than at other regions of the Domain III rRNA. Of the 33 nucleotides (nt) that report a difference in SHAPE reactivity of $\geq$40% between Domain III^{23S} and Domain III^{alone} in the presence of magnesium, 25 are seen to be involved in direct inter-domain interactions (<3.4 Å interatomic distances) in the LSU or are in close proximity to those nucleotides involved in inter-domain interactions. This pattern suggests that Domain III^{alone} folds into a near-native state, and that "insertion" of Domain III into the 23S rRNA (to form Domain III^{23S}) primarily affects the nucleotides involved in inter-domain interactions. A detailed list of other inter-domain interactions is available in Tables 7, 8 (Appendix B); the tables also indicate if SHAPE detects these interactions.

Specifically, the 3D structure of the LSU shows that A1284 forms base–backbone hydrogen bonds with G489 of Domain I. Nearby A1287 forms base–base stacking interactions with C1648 of Domain IV. As seen in Figure 13B, adding Domain III back into the 23S rRNA changes the SHAPE reactivities of A1284 and A1287. Similar changes in SHAPE reactivities are seen for (i) fragment G1325–G1332, where G1325 forms base–backbone hydrogen bonds with A1269 and C1270 of Domain II, and U1326 forms base–backbone hydrogen bonds with C1648 and G2010 of Domain IV; (ii) A1365, which forms base–backbone hydrogen bonds with G187 of Domain I; (iii) nucleotides G1568–A1570, where A1569 forms van der Waals contacts with C693 of Domain II; and (iv) nucleotides A1616–C1617, where

C1617 forms base–backbone and backbone–backbone hydrogen bonds with C749 and A750 of Domain II. This pattern of differential SHAPE reactivity indicates the subtle structural changes that occur when Domain III forms inter-domain interactions with other elements of the 23S rRNA. These inter-domain interactions are disrupted when Domain III is excised from the 23S rRNA, while the intra-domain interactions are conserved.

4.3 Discussion

The domain structures of rRNAs have profound implications for folding and function of the ribosome, and early evolution of life. In contrast to the SSU, it has been proposed that the 2° domains of the LSU (Fig. 10A) are melded into a single monolithic unit [139, 10, 167]. LSU 2° domains are thought to be so highly intertwined and interconnected that they lack distinct structural and functional significance and are not true 3D domains.

Considering the extensive network of intra-domain tertiary interactions of Domain III (Fig. 10B; Tables 5, 6, Appendix B) and its isolation from the inter-domain network of molecular interactions within the LSU, we hypothesize that Domain III is a true 3D domain. In contrast, Domain V, which contains the Peptidyl Transferase Center, is extensively networked with other 2° domains. Domain V makes 24 inter-domain A-minor interactions [17]. Additionally, Domain V makes six inter-domain magnesium-mediated phosphate–phosphate linkages [61]. Domain III only makes six A-minor interactions and single magnesium-mediated phosphate–phosphate linkage with other 2° domains.

We present data indicating that Domain III$^{\text{alone}}$ adopts a secondary structure that is the same as Domain III$^{\text{23S}}$ (Figs. 10C, 13A). The addition of magnesium facilitates folding to a near-native state of both Domain III$^{\text{alone}}$ and Domain III$^{\text{23S}}$, with the formation of intra-domain tertiary interactions (Figs. 10D, 13B). The disruption of inter-domain interactions of Domain III is reflected in the subtle but observable changes in SHAPE reactivity when Domain III is excised from the 23S rRNA (i.e., when Domain III$^{\text{23S}}$ is converted to Domain III$^{\text{alone}}$) (Fig. 13B). The mapping of these changes in SHAPE reactivity to regions of inter-domain interactions is evidence that Domain III$^{\text{alone}}$ and Domain III$^{\text{23S}}$ fold to near-native states. This interpretation is supported by the previous observation that Domain III$^{\text{alone}}$

interacts specifically with ribosomal protein L23 [110].

In sum it appears that, like the SSU, the LSU also contains some elements of a 3D domain-based architecture, in spite of its monolithic appearance. At least some 2°domains of the 23S rRNA (Domain III) autonomously fold to near-native states apart from the rest of the LSU. Consequently, at least some LSU 2° domains may have played roles similar to SSU 2° domains during the evolutionary development of the ribosome. Previous support for the importance of 3D domains of the LSU is found in the demonstration that Domain I alone is highly structured [44]. Further, Garret and colleagues have demonstrated that isolated domains of the 23S rRNA are able to form the correct secondary structure and bind to specific ribosomal proteins [44, 77, 110].

4.3.1 Evolutionary implications of the domain structure of the Domain III

The ribosome in its present form was well-established at the emergence of the last universal common ancestor of life (LUCA) [160, 162, 132, 17, 60, 11, 48]. There is a consensus that some parts of the ribosome are even older than LUCA, predating the protein world. Parts of Domain V of the 23S rRNA are believed to be among the most ancient parts of the ribosome [160, 162, 132, 17, 60, 11, 48] while Domain III is thought to be a more recent addition [63]. The data presented here support the hypothesis that Domain III was added as an intact entity to the ancestral ribosome—assuming that the 3D domain is a unit of ribosomal evolution. This evolutionary model is consistent with the absence of Domain III from certain mitochondrial rRNAs, such as that of *Trypanosoma brucei* [131]. Ribosomes in which Domain III is absent may have had this domain deleted by relatively recent evolutionary processes within the mitochondrion, but presumably retained functionality with the assistance of proteins.

4.4 Materials and Methods

T. thermophilus rRNA transcripts were produced and purified as described in Appendix B.

4.4.1 SHAPE reactions

Magnesium was removed from 25 pmol of Domain III or 23S RNA in 32 μL 1$\times$ TE (10 mM Tris-HCl, 1 mM EDTA, pH 8.0) by heating in the presence of magnesium chelating resin (Hampton Research) to 95°C for 3 min, followed by chilling on ice. Thirty-two microliters of Mg^{2+}-free RNA was mixed with 4 μL 10$\times$ folding buffer (500 mM HEPES pH 8.0, 2 M sodium acetate pH 8.0) and then incubated at 37°C for 20 min. For RNA folding with Mg^{2+}, the 10$\times$ folding buffer contained 500 mM HEPES pH 8.0, 2 M sodium acetate pH 8.0, 100 mM $MgCl_2$.

The folded RNA was divided equally between two tubes. To one tube, 2 μL of 130 mM NMIA (or 800 mM BzCN) in anhydrous DMSO was added, while the other half served as a negative control to which 2 μL pure DMSO was added. The reactions were incubated at 37°C for 1 h with NMIA. The modification reaction using BzCN is complete in a few seconds at room temperature [99]. Denaturing SHAPE experiments were performed in 20 mM HEPES pH 8.0 (final concentration) for 4 min at 90°C using 130 mM NMIA in anhydrous DMSO. The modified RNA was purified using RNeasy Mini Kit (Qiagen) and resuspended in 20 μL 1$\times$ TE. The recovery after purification was 65%–75%.

A 20-nt long DNA oligomer 5'-CGCGCCTGAGTGCTCTTGCA-3', that anneals to the 3'-end of Domain III, was used to prime the reverse transcription. The primer was labeled with 6-FAM using a 5'-amino C6 linker (Operon MWG). Twenty microliters of modified RNA was added to 8 pmol of the primer in 10 μL of 1$\times$ TE. The RNA-primer solution was heated to 95°C for 1 min and cooled to 30°C over 45 min at a rate of 1.4°C/min. After primer annealing, SuperScript III Reverse Transcriptase buffer (Invitrogen) was added at 30°C. The solution was heated to 55°C for 1 min and reverse transcription was initiated by adding 1 μL (200 U) of SuperScript III Reverse Transcriptase (Invitrogen). The reaction was incubated at 55°C for 2 h and quenched by heating to 70°C for 15 min. Di-deoxy sequencing reactions used unmodified Domain III RNA and 1 mM ddNTPs (TriLink BioTechnologies). One microliter of the reverse transcription reaction mixture was mixed with 0.3 μL ROX-labeled DNA sizing ladder and 8.7 μL of Hi-Di Formamide (Applied Biosystems) in a 96-well plate. The mixture was heated to 95°C for 5 min to denature the cDNAs and resolved on a

3130 Genetic Analyzer (Applied Biosystems) using custom fluorescence spectral calibration. Capillary electrophoresis data were processed as described in Appendix B.

4.4.2 Tertiary interactions

A detailed description of the protocol followed to annotate the intra-domain and inter-domain tertiary interactions observed for Domain III is available in Appendix B.

4.5 Acknowledgments

This work was supported by the NASA Astrobiology Institute and the Center for Ribosomal Origins and Evolution. We thank Timothy K. Lenz for helpful discussions.

CHAPTER V

IN VITRO SECONDARY STRUCTURE OF THE GENOMIC RNA OF SATELLITE TOBACCO MOSAIC VIRUS[1]

Abstract: Satellite tobacco mosaic virus (STMV) is a $T = 1$ icosahedral virus with a single-stranded RNA genome. It is widely accepted that the RNA genome plays an important structural role during assembly of the STMV virion. While the encapsidated form of the RNA has been extensively studied, less is known about the structure of the free RNA, aside from a purported tRNA-like structure at the 3' end. Here we use selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) analysis to examine the secondary structure of *in vitro* transcribed STMV RNA. The predicted secondary structure is unusual in the sense that it is highly extended, which could be significant for protecting the RNA from degradation. The SHAPE data are also consistent with the previously predicted tRNA-like fold at the 3' end of the molecule, which is also known to hinder degradation. Our data are not consistent with the secondary structure proposed for the encapsidated RNA by Schroeder et al., suggesting that, if the Schroeder structure is correct, either the RNA is packaged as it emerges from the replication complex, or the RNA undergoes extensive refolding upon encapsidation. We also consider the alternative, i.e., that the structure of the encapsidated STMV RNA might be the same as the *in vitro* structure presented here, and we examine how this structure might be organized in the virus. This possibility is not rigorously ruled out by the available data, so it remains open to examination by experiment.

[1]This chapter was adapted from Athavale, S. S., Gossett, J. J., Bowman, J. C., Hud, N. V., Williams, L. D., and Harvey, S. C., "In vitro secondary structure of the genomic RNA of satellite tobacco mosaic virus," *PLoS ONE*, vol. 8, no. 1, p. e54384, 2013.

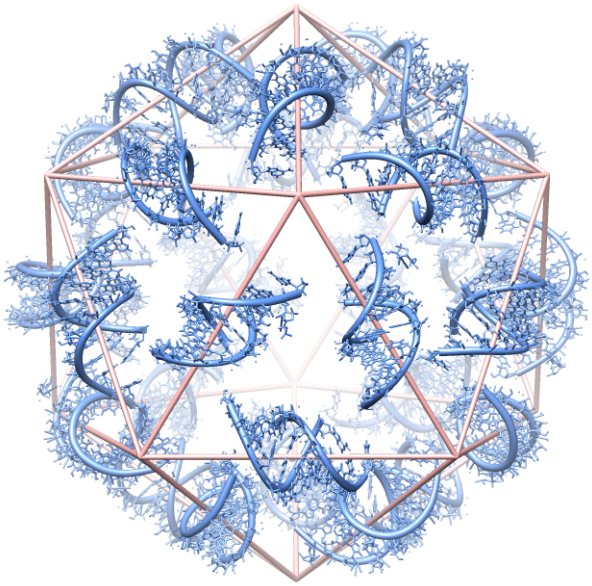


Figure 14: Distribution of double-helical RNA segments in the STMV virion. The crystal structure of STMV [70] reveals 30 segments of double-helical RNA (blue). Each helix contains 9 base pairs, centered on a crystallographic two-fold axis. An icosahedral cage (pink) is shown for reference. Adopted from [168].

5.1 Introduction

Satellite tobacco mosaic virus (STMV) is a $T = 1$ icosahedral virus with a single-stranded, positive-sense RNA genome, 1058 nucleotides in length. A capsid of 60 identical protein subunits surrounds the genome in the STMV particle. Like other satellite viruses, STMV encodes its own capsid protein but requires a helper virus for replication. For a review on the general properties of STMV, see Dodds [39]. STMV has been studied extensively as a model for the assembly of other single-stranded RNA viruses [127], and as a vector for the delivery of foreign genes into tobacco plants [52].

Efforts to characterize the RNA and its role in assembly have produced mixed results. The virus crystal structure has been solved at 1.8 Å resolution [70], although some of the protein and 41% of the RNA are not visible in the electron density maps. The RNA that is visible is revealed as 30 double-helical segments, each 9 base pairs in length and closely associated with dimers of coat protein (Figure 14). The helical axes are perpendicular to the icosahedral 2-fold axes, forming part of the edges of an icosahedron. With this constraint on the structure, Larson and McPherson proposed that the RNA forms a series of stem-loop

substructures, with only short-range (local) base pairing. They suggested that coat proteins bind to successive stem-loops as these form upon emerging from the replication complex [71]. The results of atomic force microscopy (AFM) experiments are consistent with this hypothesis [69].

Schroeder et al. used chemical probing to examine the RNA structure inside the virus. They combined these data with the assumption of co-replicational folding to produce an ensemble of models for the secondary structure [128]. Each of these contains a series of 30 stem loops, with local base pairing; it is important to emphasize that the absence of long-range base pairs is an assumption built into the model, not a hypothesis that was tested by the chemical probing. They reported a single "most representative" secondary structure from that ensemble. We recently used that secondary structure to develop an all-atom model for the mature virus [168], containing every single amino acid and every single nucleotide. (We believe this is the first such model for any virus.).

The capsid-free form of STMV RNA has been relatively overlooked in structural studies, in part because the secondary structure of the encapsidated RNA is believed to be different than the free RNA [71]. A tRNA-like structure (TLS) has been predicted at the 3' end of the molecule [46, 55], but there is no evidence in the crystallographic data for or against its existence in the encapsidated RNA. A feature seen in AFM images of phenol extracted RNA could be interpreted as the predicted TLS [69], but Schroeder et al. have concluded that the TLS is not compatible with their chemical modification data [128]. Larson et al. have argued that, if the tRNA-like structure and replication recognition site structure were maintained inside the virus, there would be insufficient RNA remaining to connect the stem-loop segments [70].

Here we report a secondary structure model for *in vitro* transcribed STMV RNA, based on chemical probing data obtained using selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) [93]. SHAPE provides information on local nucleotide dynamics [92], thus reflecting the extent to which a nucleotide is constrained by base pairing or other interactions [38]. The SHAPE signal is highly correlated with Watson-Crick base pairing [15], and is capable of significantly improving the accuracy of RNA secondary structure

predictions [38]. Our primary motivation for this work is to establish the secondary structure for the free STMV RNA, in the absence of the capsid protein. We also compare our probing data to the secondary structure proposed by Schroeder et al. for the RNA *in virio*, [128], and to the predicted tRNA-like structure at the 3' end of the RNA [46, 55].

5.2 Results and Discussion

5.2.1 SHAPE Analysis of the Free form of STMV RNA

SHAPE [93] involves treating the RNA with an electrophilic reagent that reacts selectively at the ribose 2'-OH position of conformationally flexible nucleotides to form 2'-*O*-adducts. Reverse transcription using fluorescently labeled primers gives cDNA fragments whose lengths are determined by locations of the 2'-*O*-adducts, and whose quantities can be measured by capillary electrophoresis.

We first probed the *in vitro* transcribed STMV RNA in the presence of 250 mM Na^+ using the SHAPE reagent N-methylisatoic anhydride (NMIA). Under these conditions (no Mg^{2+}), one would expect the formation of secondary structure, but not necessarily tertiary structure [22, 42, 19]. We obtained good quality SHAPE reactivity data on 1029 nucleotides, or 97.3% of the 1058-base long STMV RNA. Nucleotides 1–4 and 1034–1058 were omitted from the analysis. The normalized SHAPE reactivity values for STMV RNA ranged from -0.17 to 2.34 with the exception of nucleotide 427, whose reactivity was an outlier at 7.25. Nucleotides with normalized reactivity values <0.3 are considered unreactive; 0.3 to 0.7, moderately reactive; >0.7, highly reactive [38]. Using these criteria, we observed 727 unreactive nucleotides, 189 moderately reactive nucleotides, and 113 highly reactive nucleotides. Six nucleotides—244, 427, 449, 469, 887, and 974—also met the criterion for hyper-reactivity, i.e., normalized reactivity >2 [92]. The data processing procedures are given in more detail in the Methods section, and in Appendix C.

5.2.2 The SHAPE-restrained STMV RNA Secondary Structure Contains Long-range Base Pairing

The SHAPE reactivity information was incorporated into the thermodynamic folding algorithm RNAstructure [118] as a pseudo-free energy change term [38] to predict a secondary

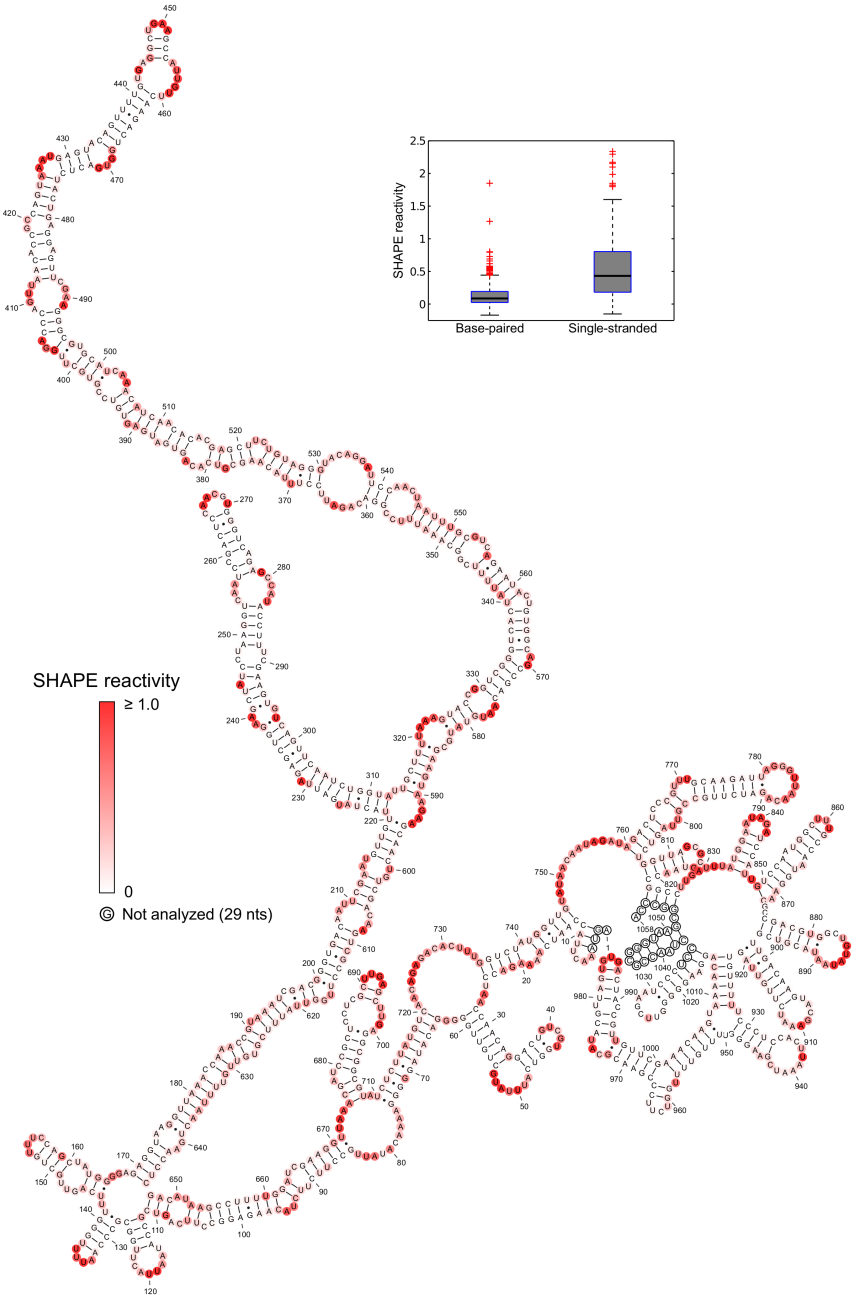


Figure 15: SHAPE-restrained secondary structure model for free STMV RNA. Nucleotides are colored according to their SHAPE reactivity (see scale). Inset shows a box plot comparing the distribution of SHAPE reactivity values between base paired and single-stranded nucleotides. Each grey box represents the interquartile range (IQR) of the data; the bottom and top edges of the box are the 25th and 75th percentiles, respectively. The band near the middle of each box is the median value. The whiskers above and below each box extend to the most extreme data points not considered outliers. Outliers are plotted individually as crosses. Points are outliers if they are greater than 1.5 IQR from the 75th percentile or less than 1.5 IQR from the 25th percentile. In this secondary structure model, the distribution for base paired nucleotides is narrower and has a much lower median value than the distribution for single-stranded nucleotides.

structure model for the free STMV RNA (Figure 15). In the virus, it has been proposed that there are 30 stem-loops [70]. This proposal was incorporated into the Schroeder model by prohibiting long-range base pairing [128]. We imposed no restriction on the distance along the primary sequence between base-paired nucleotides, since there is no *a priori* reason for doing so for an RNA probed *in vitro*.

We recognize that chemical probing cannot define a single secondary structure [67, 32], because SHAPE reactivity is inversely correlated with base pairing, but the correlation is not perfect; some base paired nucleotides are reactive, and some unpaired nucleotides are not. To address this issue, we report the structure that is most consistent with the SHAPE data (Figure 15), along with several suboptimal structures (Figure 34, Appendix C), also generated by RNAstructure.

We evaluated the agreement between the model and the data by comparing the distribution of reactivity values in single-stranded nucleotides with the distribution of reactivities in base paired nucleotides (Figure 15, inset box plot). The reactivities of base paired nucleotides are less disperse and have a much lower median value than the reactivities of single-stranded nucleotides. These distributions are consistent with SHAPE experiments on RNAs with known secondary structures [156].

The SHAPE-restrained secondary structure is characterized by significant long-range base pairing and minimal branching, especially for the region between nucleotides 169 and 646. This region, consisting of double-helical segments broken intermittently by small internal loops and bulges, is reminiscent of *in vitro* transcribed viroid RNA [163]. The SHAPE-restrained structure is noticeably different from the minimum free energy (MFE) structure (Figure 16) predicted using RNAstructure [118]. Unsurprisingly, the MFE structure is less consistent with the SHAPE data.

5.2.3 Maximum Ladder Distance of the SHAPE-restrained STMV RNA Secondary Structure is Much Larger than Expected

The SHAPE-restrained secondary structure of STMV RNA appears unusually highly extended. To evaluate the extendedness of this secondary structure, we used a metric first introduced by Yoffe et al. [165], the maximum ladder distance (MLD). MLD is the largest

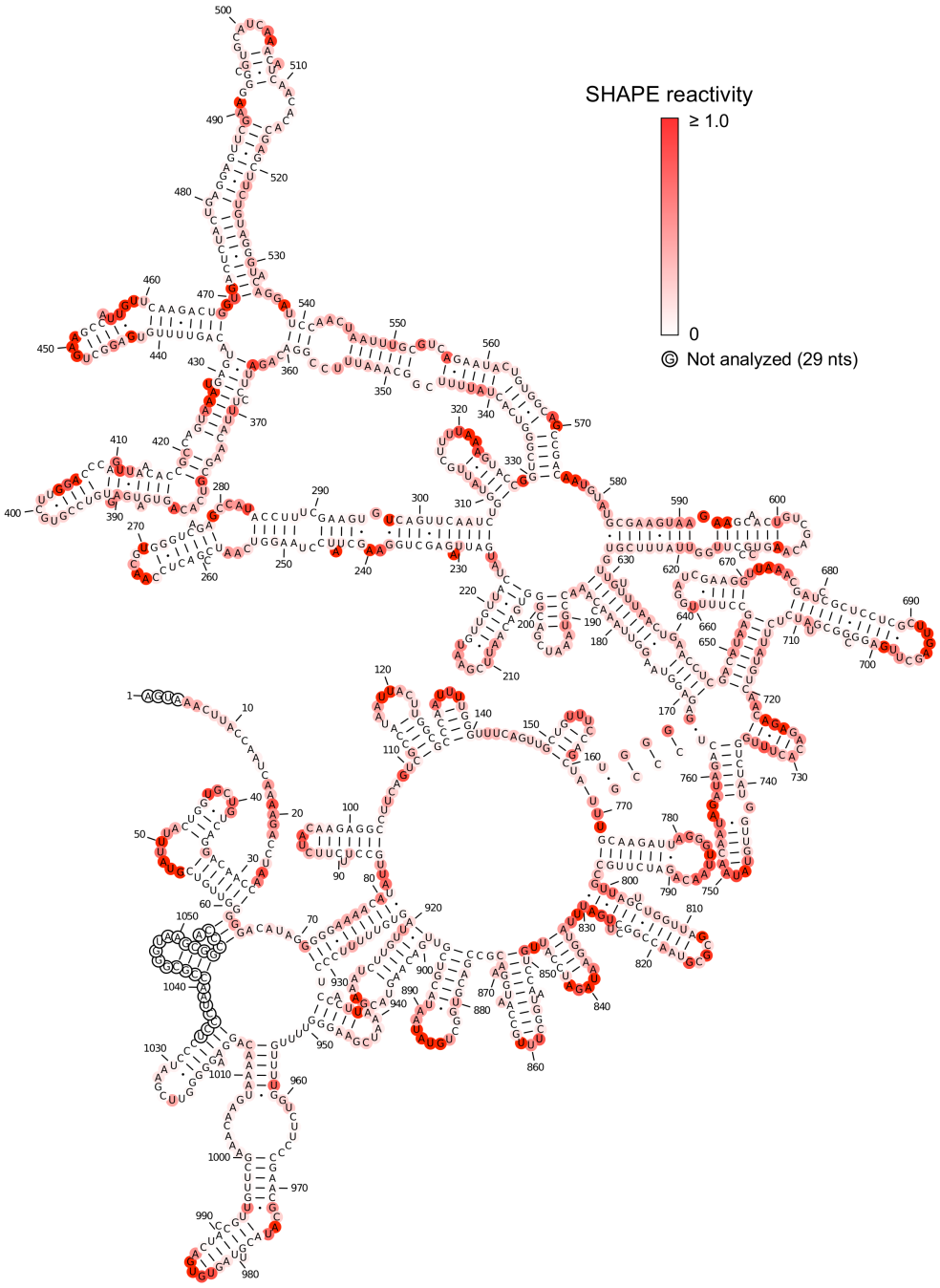


Figure 16: Minimum free energy (MFE) structure obtained for STMV RNA without the SHAPE data. The structure was predicted using RNAstructure with default parameters. Nucleotides are colored according to their SHAPE reactivity (see scale). The SHAPE data are not consistent with this model, since several base paired regions have high reactivity values.

value of ladder distance, LD_{ij}, for all combinations of i and j, where LD_{ij} is the number of base pairs that are crossed along the most direct path from base i to base j in the standard two-dimensional graph representing the secondary structure. Yoffe et al. previously used this measure to compare RNAs of $T = 3$ icosahedral viruses with a set of random RNA sequences with virus-like compositions [165]. For a given RNA sequence, they generated an ensemble of secondary structures, calculated the MLD for each of these and reported the average, designated $\langle MLD \rangle$. As a control, they generated an ensemble of secondary structures from shuffled sequences and calculated the $\langle MLD \rangle$ for that ensemble. They found that the RNA genomes of self-assembling icosahedral viruses have smaller $\langle MLD \rangle$ values than do shuffled sequences, i.e., viral RNA secondary structures are predicted to be more highly branched than those of random sequences. They suggested that these viral RNAs would therefore have compact three-dimensional structures, facilitating viral assembly.

The MLD of the SHAPE-restrained secondary structure (Figure 15) is 205. For comparison, the MLD of the more branched MFE structure (Figure 16) is 101, while $\langle MLD \rangle = 146.7$ for a collection of 1000 suboptimal structures. Remarkably, the experimental MLD is higher than the MLD of any of the suboptimal structures (Figure 17, top panel). We estimated the probability distribution for MLD values of random RNAs with the same length and nucleotide composition as STMV (Figure 17, bottom panel), finding that it is highly unlikely that a secondary structure with an MLD this high would have occurred by chance ($P < 0.004$).

We have also examined the MLDs of a series of suboptimal SHAPE-restrained structures, generated by RNAstructure (Figure 34, Appendix C). The first five of these all have similar, highly elongated structures, with MLDs of 169 or greater; the pseudoenergies of these structures range from -798 kcal/mol for the structure in Figure 15, to -784 kcal/mol for the fifth suboptimal structure. Structures with shorter MLDs (≤ 124) all have higher pseudoenergies (-770 kcal/mol or above), so they are clearly inconsistent with the SHAPE data.

This model of the STMV RNA secondary structure is at variance with the observation of Yoffe et al. that RNAs of small icosahedral viruses have smaller MLDs than do random

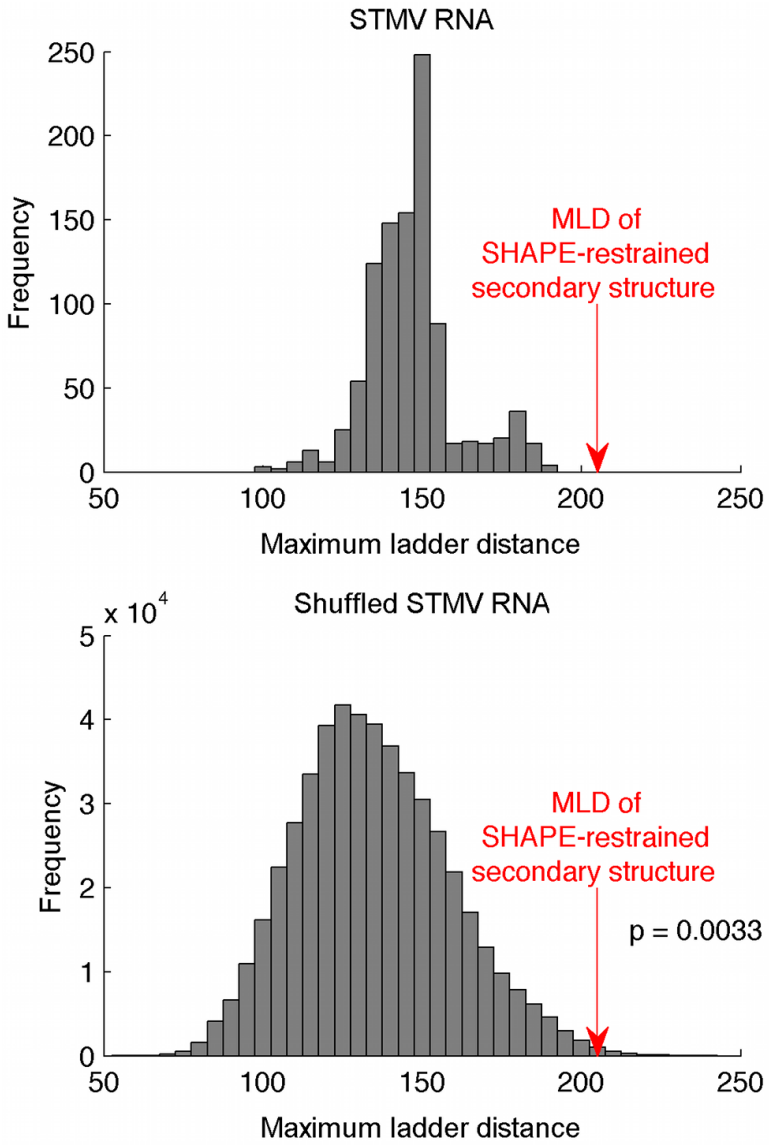


Figure 17: Histogram of maximum ladder distance values calculated for STMV RNA and shuffled STMV RNA sequences. The MLD of the SHAPE-restrained structure is much higher than the MLDs of 1000 suboptimal structures predicted for the STMV RNA sequence (top). The extreme MLD of the SHAPE-restrained structure is unlikely to have occurred by chance: the bottom histogram was obtained using 1000 suboptimal structures for each of 500 randomly shuffled sequences with the same length and nucleotide composition as STMV. Fewer than 0.4% of these structures have MLDs greater than the MLD of the SHAPE-restrained STMV structure.

sequences. We note, however, that their observations were based on data for $T = 3$ viruses with RNA genomes with lengths greater than 2500 nucleotides, while STMV is a $T = 1$ virus with a much smaller genome. Furthermore, it has been argued that STMV assembles as the RNA is replicated [71]. If so, then the $\langle MLD \rangle$ of STMV RNA is not relevant for assembly, since the RNA would not be in thermodynamic equilibrium, an implicit assumption made by Yoffe et al.

5.2.4 SHAPE Probing Supports a tRNA-like Structure (TLS) at the 3' End of STMV RNA

The 240 3'-terminal nucleotides of STMV RNA have more than 65% overall sequence similarity with the corresponding nucleotides of TMV U1 RNA, including two nearly identical regions of approximately 40–50 bases each [96]. On the basis of phylogenetic comparisons, Felden et al. proposed that the 3' end of STMV RNA folds into a tRNA-like structure similar to that found in TMV RNA [46]. The authors also demonstrated that the STMV RNA could be aminoacylated *in vitro* with histidine, although STMV RNA charging is less efficient than TMV RNA.

In a related study, Gultyaev et al. predicted a secondary structure for the 406 3'-terminal nucleotides of STMV RNA [55]. In addition to a tRNA-like structure at nucleotides 873–1058, their model included a stretch of three consecutive pseudoknots at nucleotides 653–727 and five stem-loops at nucleotides 735–870. Our SHAPE data support the predicted tRNA-like structure and the five stem-loops, but they are mostly inconsistent with the predicted pseudoknots at nucleotides 653–727 (Figure 18). It is important to note that the last 25 nucleotides at the 3' end are missing in our analysis due to experimental limitations.

Since the RNAstructure program does not allow pseudoknots in its calculations, the tRNA-like structure and associated pseudoknots would not show up in any SHAPE-restrained secondary structure prediction of STMV RNA. Therefore, we built an alternate model of the genome by combining the SHAPE-restrained secondary structure predicted separately for nucleotides 1–727 with the Gultyaev prediction for nucleotides 728–1058 (Figure 19). This produces structures for the 5' and 3' ends of the RNA that

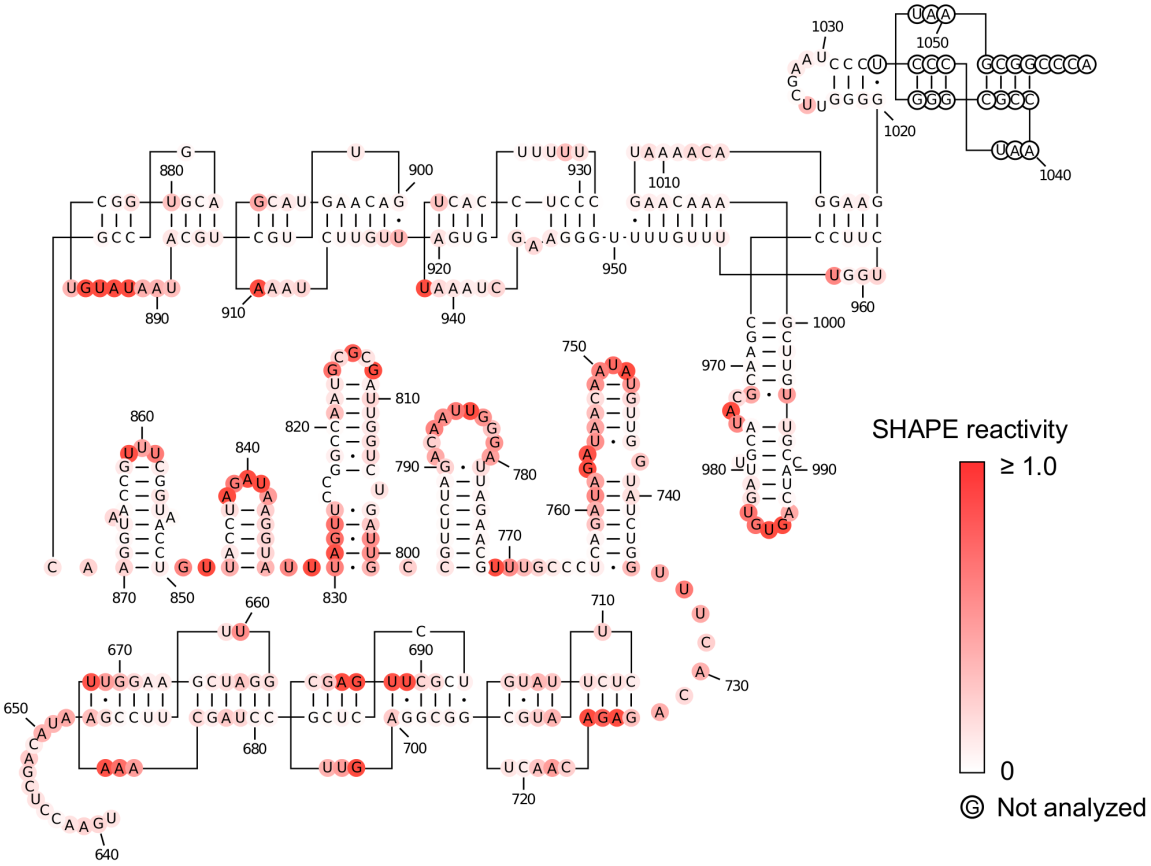


Figure 18: Predicted secondary structure at the 3' end of STMV RNA. Secondary structure for the 406 3'-terminal nucleotides of STMV RNA proposed by Gultyaev et al. [55]. Nucleotides are colored according to their SHAPE reactivity (see scale). The SHAPE data supports the tRNA-like structure and the five stem-loops (nucleotides 728–1058), but does not support the second pseudoknot domain (nucleotides 653–727).

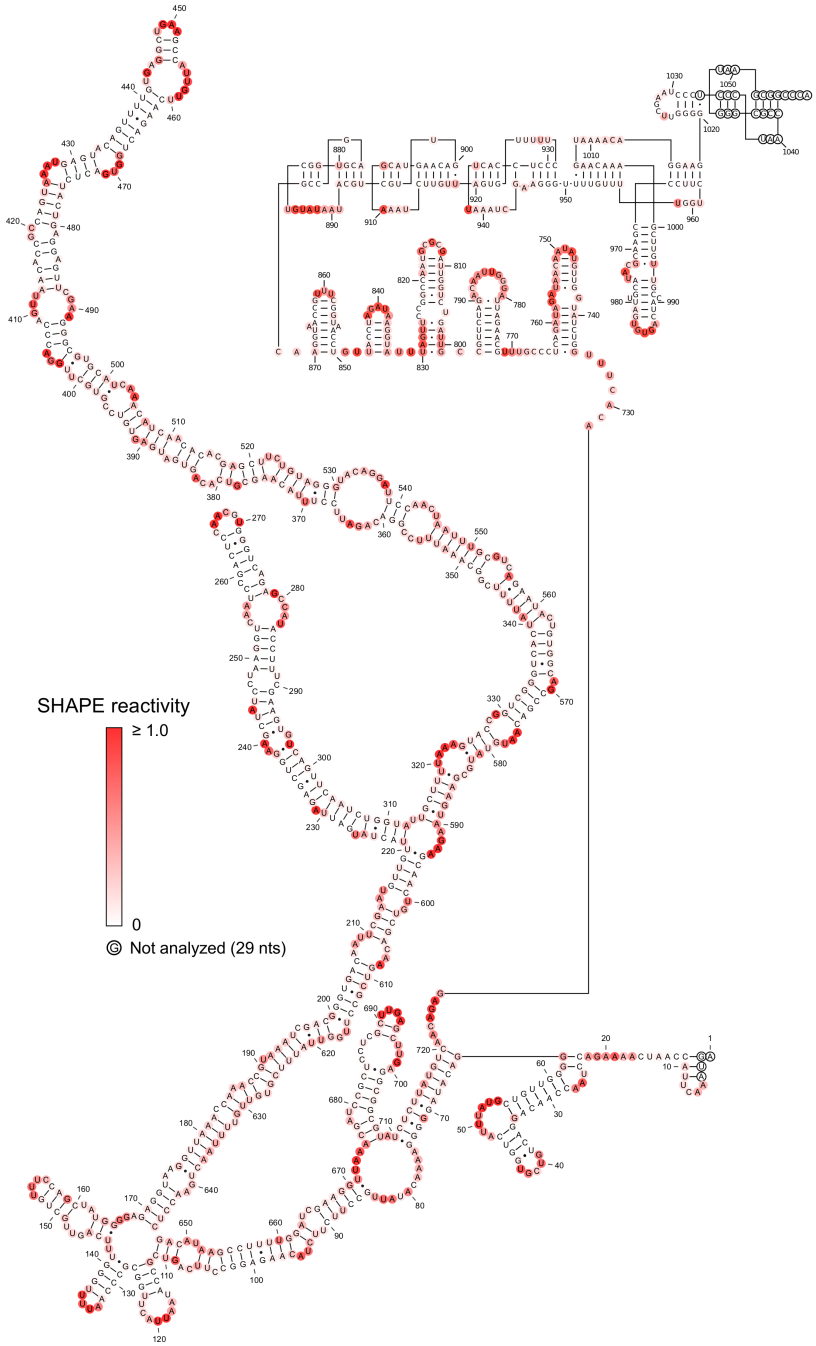


Figure 19: SHAPE-restrained secondary structure of free STMV RNA with a tRNA-like fold at the 3' end. This alternate model of the STMV RNA was obtained by combining the SHAPE-restrained secondary structure predicted separately for nucleotides 1–727 (Figure 15) with the Gultyaev et al. prediction [55] for nucleotides 728–1058 (Figure 18). Nucleotides are colored according to their SHAPE reactivity (see scale). The extended central domain (nucleotides 64–720) is identical to that of Figure 15.

differ from the structure shown in Figure 15, but the very extended central domain (nucleotides 64–720) is identical to that of Figure 15. We favor the model that includes the TLS (Figure 19) over the structure in Figure 15, because of the biochemical data [46].

5.2.5 Comparison of Probing Data on Free RNA with Data on Encapsidated RNA

We compared our SHAPE reactivity data obtained on *in vitro* transcribed RNA with the Schroeder et al. chemical probing data obtained on encapsidated RNA [128]. They reported the top 161 nucleotides modified with dimethyl sulfate (DMS), carbodiimide (CMCT), or kethoxal. Of these strongly modified nucleotides, 86 were unreactive to the SHAPE reagent, 42 were moderately reactive, and 33 were highly reactive. Although this seems like a significant amount of disagreement, SHAPE probing does not always completely agree with traditional base-reactive reagents such as DMS [93, 32]. Schroeder et al. tried SHAPE probing of the STMV RNA *in virio*, finding that the signal:noise ratio was significantly lower with this reagent than with DMS, CMCT and kethoxal; they attributed this in part to the lack of a quenching step for SHAPE probing, arguing that the SHAPE reagents probably continue to react with the RNA during extraction of the RNA from the viral particle. (See Supporting Information in reference [128].).

Second, we compared our SHAPE data with the Schroeder model, finding that the agreement is not very good. In particular, Schroeder's hairpins 1, 3, 10–13, 17, 21–22, and 25 are not consistent with the SHAPE data (Figure 20). This suggests that the secondary structure of the free RNA is different than the Schroeder model for the encapsidated RNA, as previously suggested [71]. Nor is this surprising: the Schroeder structure would not be stable in solution, as it has a very high folding free energy (−83 kcal/mol) relative to either the thermodynamic minimum free energy structure in Figure 16 (−331 kcal/mol) or the SHAPE-optimized structure in Figure 15 (−309 kcal/mol). When the RNA is packaged into the virus, if it must refold to this higher energy state, the cost would presumably be paid by favorable RNA-protein interactions.

As a separate comparison, we asked whether or not the probing data of Schroeder et al. are consistent with the SHAPE-restrained model. (We are curious about the possibility

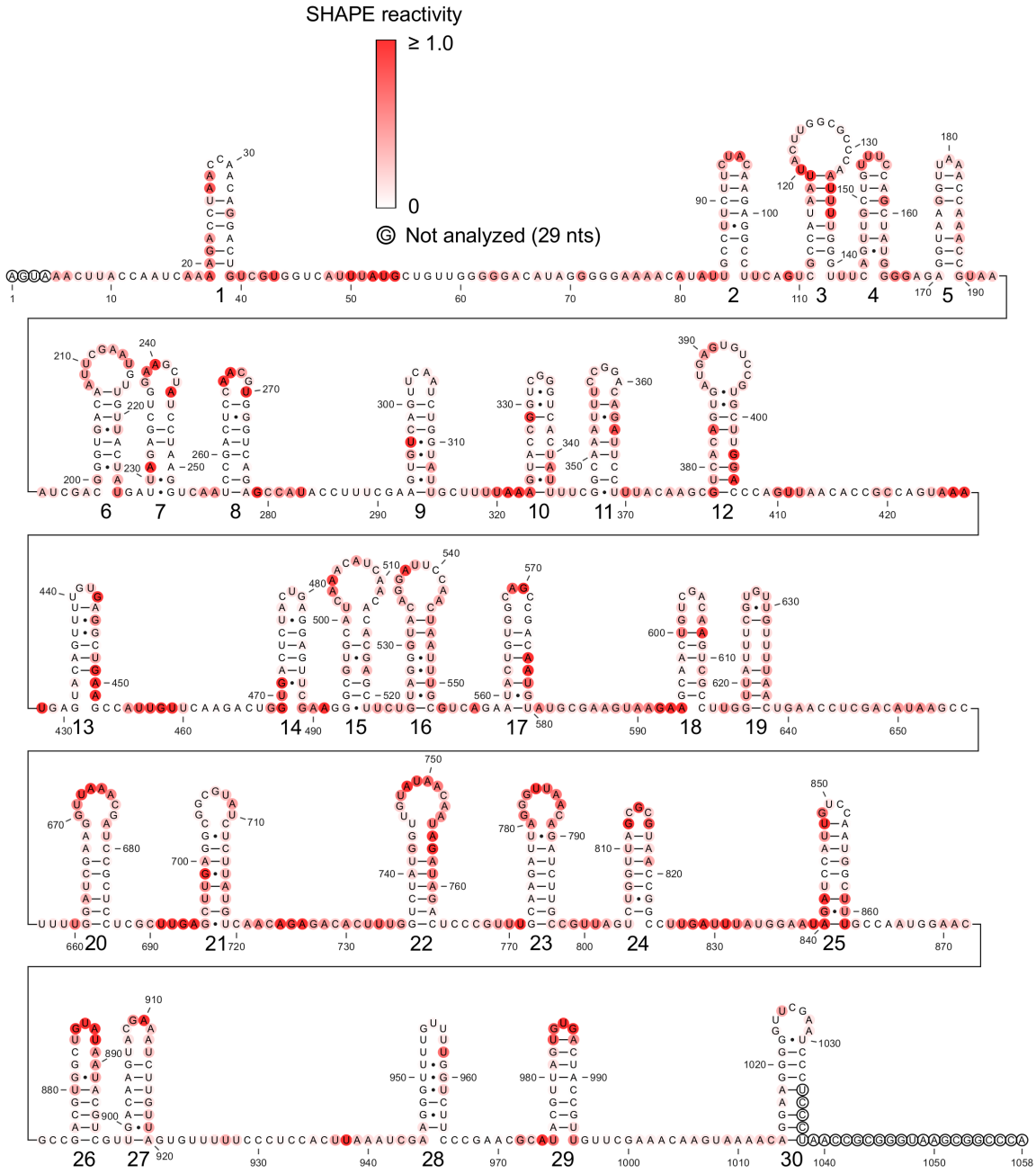


Figure 20: Schroeder secondary structure model for encapsidated STMV RNA. Schroeder et al. predicted this secondary structure on the basis of the co-replicational folding and assembly hypothesis, along with chemical probing data [128]. Nucleotides are colored according to their SHAPE reactivity (see scale), and the hairpin loops are numbered from 1 to 30. Hairpins 1, 3, 10–13, 17, 21–22, and 25 are clearly inconsistent with the SHAPE data.

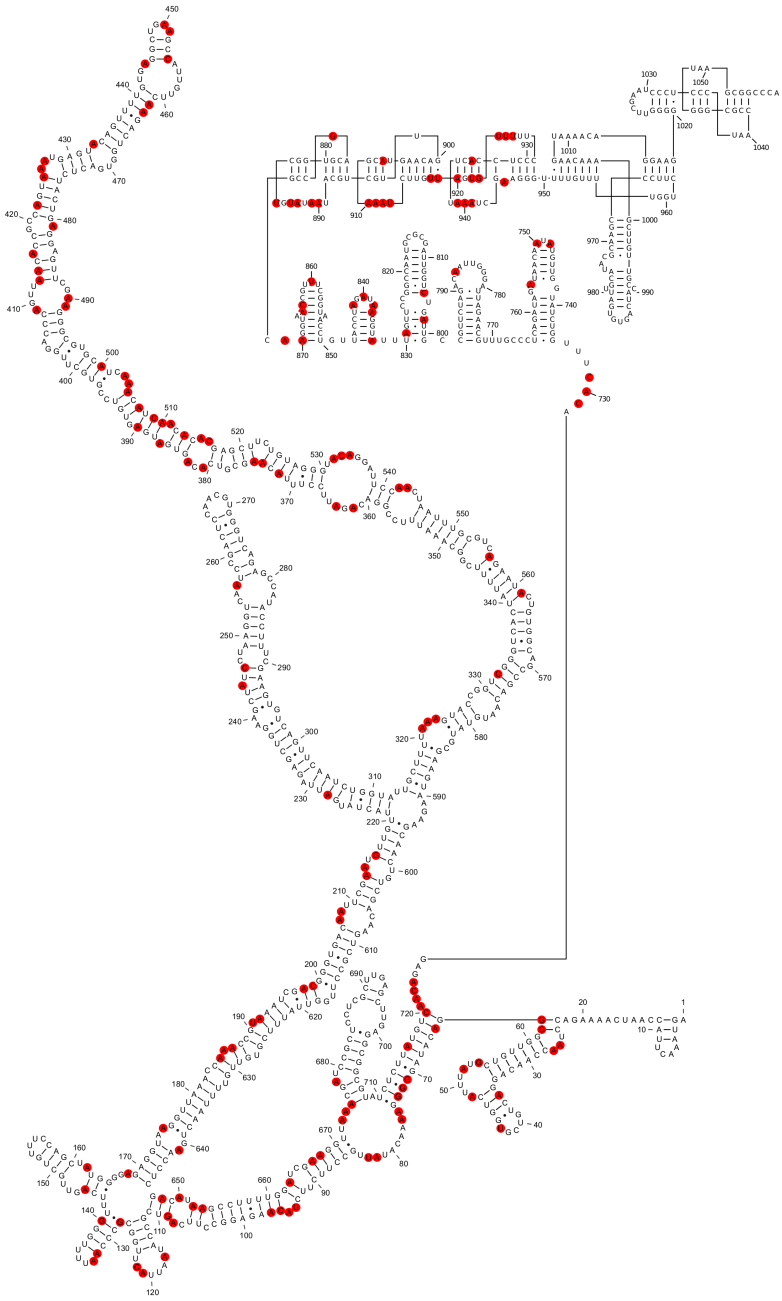


Figure 21: Mapping the chemical probing data from Schroeder et al. [128] onto the SHAPE-restrained secondary structure of *in vitro* transcribed STMV RNA. Red circles indicate nucleotides modified by DMS, kethoxal, or CMCT. The data do not appear to clearly rule out the proposed secondary structure of residues 1–730. A substantial number of the modifications occur in predicted loops, bulges, and single-stranded regions (67 out of 119 hits). Many of the reactive base-paired nucleotides are in A–U or G–U base pairs immediately adjacent to a predicted bulge loop (e.g., 128, 185, 187, 192, 213, 413–414, 556, 561, 652–653, 663, 675), while others (382–390 and 503–515) are in a predicted stem that has two bulges and has no run of more than three consecutive base pairs, so it should be prone to fraying.

that the encapsidated structure might resemble our model.) It is not possible to make a rigorous comparison, because Schroeder's data were obtained on the RNA in the mature virus, while our model represents the RNA free in solution. It is hard to evaluate how much the capsid might protect the RNA, and impossible to know which residues might be affected. It is also unclear to what extent encapsidation of a structure like ours might cause local structural disruptions. There appears to be a not unreasonable agreement between the Schroeder data and our model in the extended region (residues 1–730), and in the tRNA-like domain (Figure 21). In the extended region, the biggest disagreements lie in the stem composed of residues 384–394 and 505–514, although this is a weak stem, containing three shorter stems of only three base pairs each, separated by bulges. Otherwise, many of the hits lie in proposed bulges, or in A-U base pairs immediately adjacent to bulges. We are unable to reach a firm conclusion about what, if anything, the Schroeder data say about the possibility that this structure—or parts of it—are found in the mature virus.

5.2.6 SHAPE reactivity data for free STMV RNA with and without Mg^{2+} are not significantly different

To examine the effect of Mg^{2+} on the folding of STMV RNA, we performed an otherwise identical SHAPE experiment on the RNA in the presence of 10 mM Mg^{2+}. The presence of Mg^{2+} did not significantly change the SHAPE reactivity profile (Figure 22), indicating that STMV RNA folding is not dependent on Mg^{2+}. Some RNAs, e.g., tRNA, RNase P, the *Tetrahymena thermophila* group I intron P4-P6 domain, and domain III of the *T. thermophilus* 23S rRNA, show significant Mg^{2+}-dependence of SHAPE reactivities [143, 5, 154, 99, 4]. STMV RNA is essentially an mRNA, so its folding is not necessarily expected to be dependent on Mg^{2+}. Atomic force microscopy (AFM) images showed that STMV RNA that has been phenol-extracted from the intact virus exists in two temperature-dependent and reversible conformations, an open and a closed conformation [69]. Those authors suggested that secondary structure and significant tertiary interactions are maintained even at elevated temperature (65°C). Our SHAPE probing at 37°C suggests that either there are no significant tertiary interactions or, if there are, Mg^{2+} is not required for their formation.

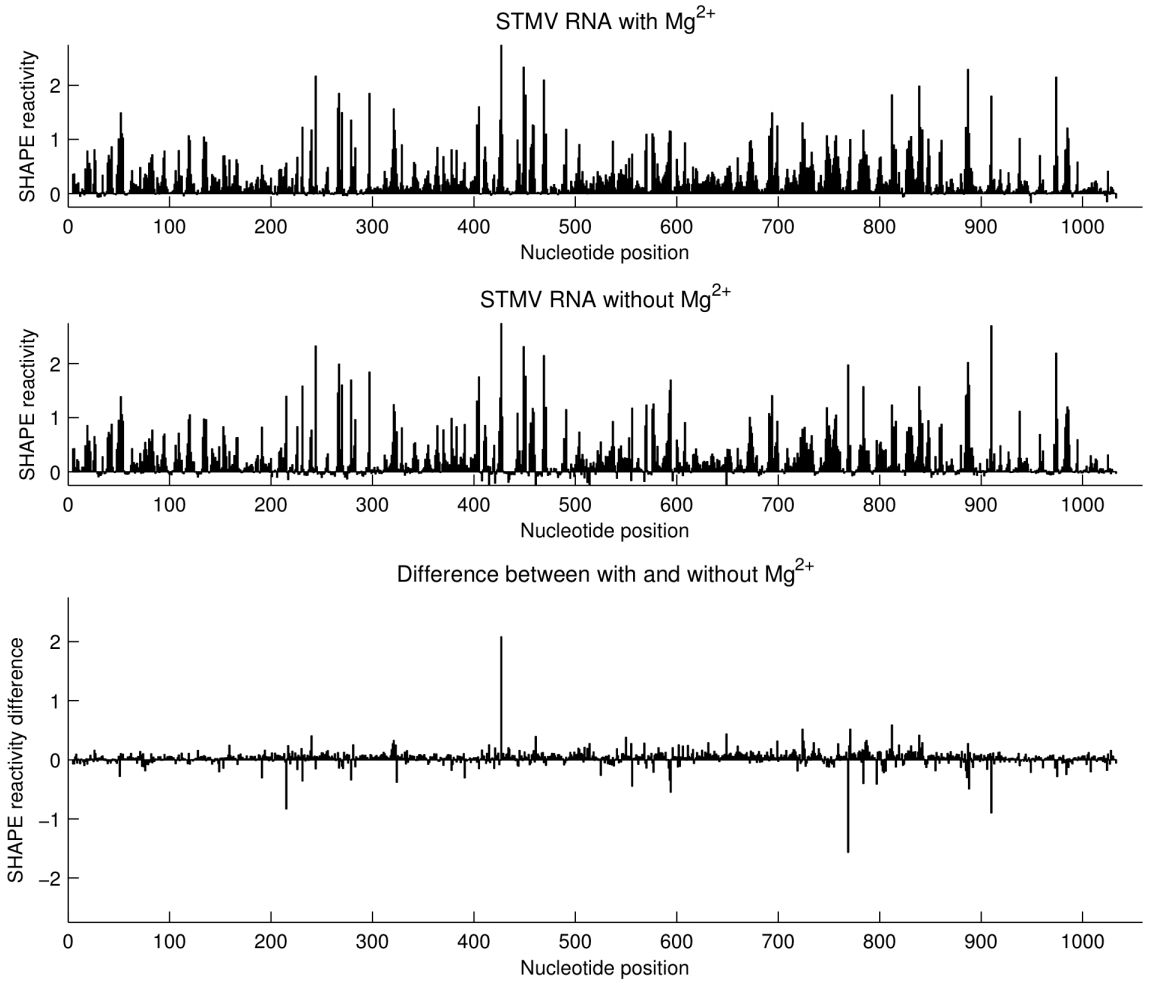


Figure 22: Effect of Mg^{2+} on the SHAPE reactivity profile of free STMV RNA. SHAPE reactivities for STMV RNA in the presence (top) and absence (middle) of Mg^{2+}. The difference plot (bottom) shows that 10 mM Mg^{2+} has little effect on the SHAPE reactivity profile.

5.2.7 Biological significance

The secondary structure proposed here (Figure 19) raises four questions.

First, is the structure of the *in vitro* transcribed RNA biologically relevant? A study by Mirkov et al. suggests that it is. They demonstrated that STMV RNA transcribed *in vitro* was biologically active, showing a consistent ability to infect tobacco plants also infected by TMV [95]. It is worth mentioning that STMV RNA can move systemically through a plant in both encapsidated and non-encapsidated forms [39, 120].

Second, does this structure play a role in viral assembly? It appears likely that the TLS represents a recognition signal for replication [46]. Also, the TLS at the 3' end of brome mosaic virus (BMV) RNA has been shown to mediate icosahedral viral assembly and function as a simple telomere [30, 43, 117, 133]. The STMV TLS might do the same.

Third, if this secondary structure is not that of the packaged RNA in the mature STMV virion, then what is its function? One plausible explanation is that it protects the RNA from degradation. Felden et al. have proposed that the tRNA-like structure (TLS) in STMV is essential for stability of its RNA [46], as has been demonstrated for TMV [49]. In addition, viroid RNAs (which are not encapsidated) have extended secondary structures, not unlike the extended domain in Figure 19. Wang et al. showed that "viroid and satellite RNAs are significantly resistant to RNA silencing-mediated degradation, suggesting that RNA silencing is an important selection pressure shaping the evolution of the secondary structures of these pathogens" [147]. This might well be the case for the extended domain of STMV RNA.

Finally, is it possible that this secondary structure is maintained inside the intact virion? As argued above, the chemical probing data from Schroeder et al. don't give a firm answer to this question. Could the extended domain be arranged to cover the edges of the icosahedron, perhaps surrounding the tRNA-like structure in the core? Figure 23 shows how our model might be organized to provide a sufficient number of double-helical stems to do this.

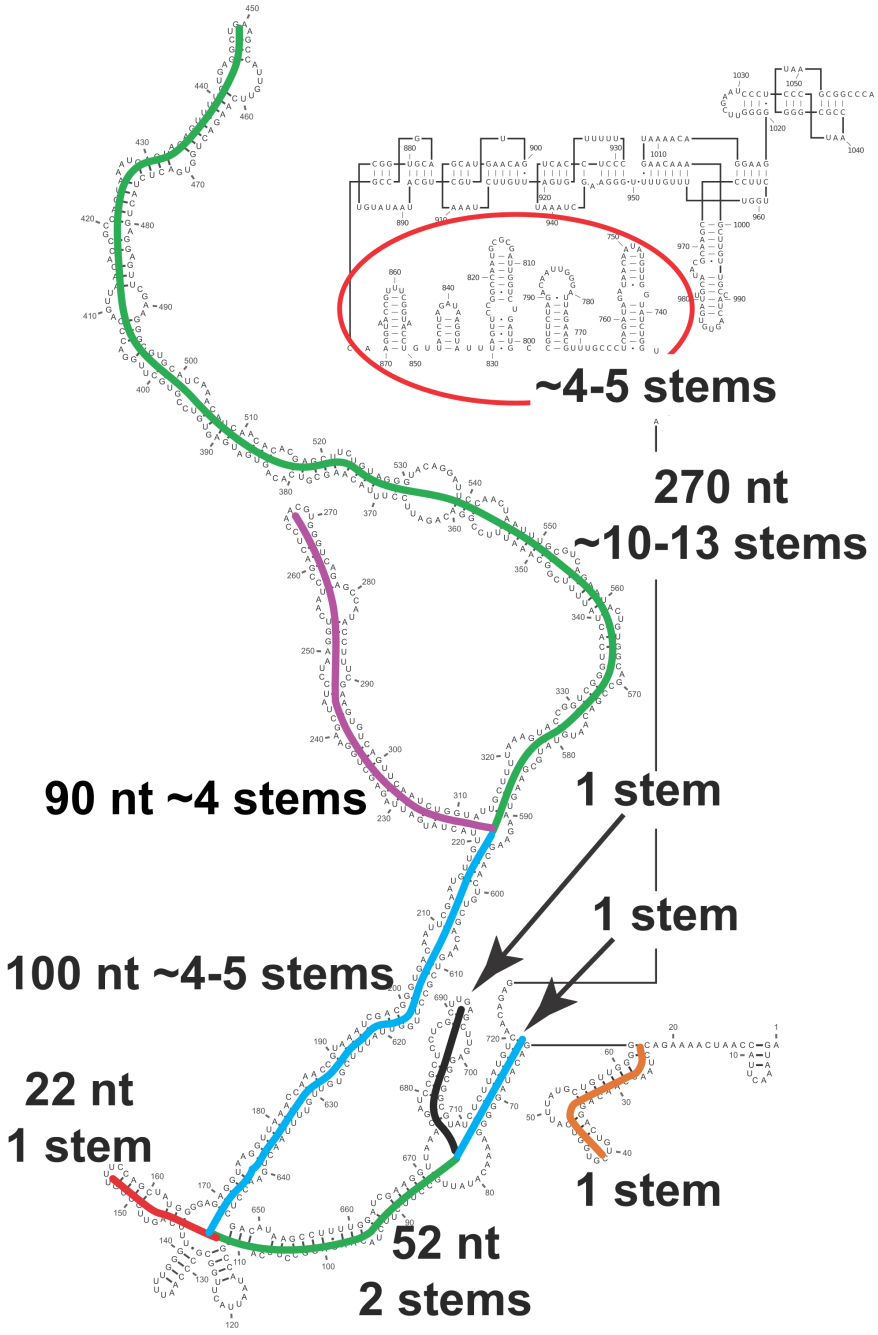


Figure 23: Identification of possible double-helical stems corresponding to those seen in the crystal structure. There are 30 stems in the crystal structure, each containing nine base pairs with an additional base stacked at each 3' end, i.e., 20 nucleotides (Figure 14). A model that connects successive stems would require something on the order of 5–10 nucleotides per connection. This figure shows how our secondary structure model might be organized to fit into the STMV capsid, with a sufficient number of stems to cover the 30 edges of the icosahedral frame, as required by Figure 14.

5.2.8 Conclusions

The SHAPE-restrained secondary structure of *in vitro* transcribed STMV RNA is highly extended, and the data support the predicted tRNA-like fold at the 3' end of the RNA [46, 55]. Both of these features may stabilize the non-encapsidated RNA *in vivo*. The predicted secondary structure of the RNA transcribed *in vitro* is considerably different from that proposed for the genome in the intact virion [128]; we have previously developed an all-atom model of the mature virus based on the latter secondary structure [168]. Here we have suggested that it might also be possible to develop a model of the mature virus using the RNA secondary structure revealed by SHAPE probing, which corresponds to the equilibrium structure.

If the genomic RNA is packaged co-replicationally, as originally proposed [71], then the Schroeder secondary structure model [128] is probably correct. Alternatively, the RNA might be fully synthesized before packaging, achieving the structure that we have proposed (Figure 19). If this is the case, then either the RNA is packaged with our structure, or it undergoes extensive refolding to achieve the Schroeder structure. Additional experimental work is needed to determine the relationship between replication and packaging, and to identify the final structure of the viral genome after packaging into STMV.

5.3 Methods

5.3.1 Preparation of STMV RNA

STMV DNA appended with a 5' T7 promoter and 3' HindIII recognition sequence was synthesized by MWG Operon and provided in a pCR 2.1-TOPO plasmid. The plasmid was cleaved with PstI (New England Biolabs), gel purified, and religated to remove an extraneous T7 promoter. The plasmid was amplified in dH5 Escherichia coli, purified using the Endo-Free Plasmid Maxi kit (Qiagen), and sequenced bi-directionally (MWG Operon). This *in vitro* transcript runs as a single band in native gel electrophoresis (Figure 35, Appendix C), suggesting a single dominant conformation.

Transcription reactions were performed by the run-off method [124], using the MEGAscript High Yield Transcription Kit (Applied Biosystems). Plasmid containing the

STMV gene was linearized with HindIII (New England Biolabs) and purified by DNA Clean & Concentrator Kit (Zymo Research). Linearized plasmid (∼0.5 μg) was transcribed in 20 μL reaction volumes for 2.5 hours at 37°C. RNA products from transcription reactions were recovered by ammonium acetate precipitation and resuspended in nuclease-free water (IDT). Yields were quantified by UV absorbance and purity by denaturing PAGE.

5.3.2 SHAPE Probing of STMV RNA

SHAPE probing of STMV RNA was performed as described in [4]. Five 20-nt long DNA primers were used to primer reverse transcription reactions. The primers were labeled with 6-FAM at the 5' end (Eurofins MWG Operon). The primers were named according to the most 5' nucleotide of STMV RNA to which they anneal: 201, 5'-ACAACATTCGAATTGTC ACC-3'; 411, 5'-TCATTTACTGGCGGTGTTAA-3'; 668, 5'-AGGAGCGGATCGTTTAAC CT-3'; 831, 5'-ACAATGGATCTATTCCATAA-3'; and 1039, 5'-TGGGCCGCTTACCCGC GGTT-3'.

5.3.3 SHAPE Data Processing

We converted the capillary electrophoresis (CE) data traces, or electropherograms, into SHAPE reactivities using in-house Matlab code. This procedure has been described in detail in Athavale et al. [4]. Briefly, this involved (1) aligning the traces to one another, (2) calculating and subtracting the baseline, (3) locating the peaks, (4) quantifying the area of each peak, (5) correcting for signal decay, (6) subtracting the background, and (7) normalizing. We used a new technique to correct for signal decay (see Appendix C for details).

For the SHAPE data acquired on the RNA in 250 mM Na^+ (no Mg^{2+}), the final reactivity values represent the average of nine separate datasets: three at a concentration of 3.25 mM NMIA, three at 6.5 mM NMIA, and three at 13 mM NMIA. For the SHAPE data acquired in 250 mM Na^+ and 10 mM Mg^{2+}, the final reactivity values represent the average of three separate datasets: one at a concentration of 3.25 mM NMIA, one at 6.5 mM NMIA, and one at 13 mM NMIA. As reported earlier [5], we have validated our methods by doing SHAPE experiments on the P4-P6 domain from the *Tetrahymena* Group I ribozyme,

getting results that are similar to previous reports on the same molecule [143].

5.3.4 RNA Secondary Structure Prediction

We folded the entire STMV RNA sequence (1058 nucleotides) using the thermodynamics-based free energy minimization algorithm in the RNAstructure software package, version 5.3 [118]. For the minimum free energy (MFE) structure, we used the default parameters. When calculating the SHAPE-restrained structure, we used the '-sh' option to incorporate the SHAPE reactivities into the algorithm as restraints [38, 89], with default values for the SHAPE slope (2.6 kcal/mol) and SHAPE intercept (-0.8 kcal/mol). (We note that, since SHAPE reactivity penalizes single-strandedness for reactive nucleotides but does not absolutely prohibit base pairing, the SHAPE penalty is properly a restraint, rather than a constraint.).

5.3.5 Maximum Ladder Distance Calculations

We calculated the MLD values using a C program (provided by Aron Yoffe and co-workers, UCLA). To compute the ensemble-average maximum ladder distance ($\langle MLD \rangle$), we first generated a random sample of 1000 suboptimal structures, drawn with probabilities equal to their Boltzmann weights, using RNAsubopt, a program in the Vienna RNA software package, version 2.0 [58]. We then calculated the $\langle MLD \rangle$ as

$$\langle MLD \rangle = \sum_{i=1}^{1000} \frac{MLD_i}{1000}. \tag{3}$$

5.4 Acknowledgments

We thank Yingying Zeng, Roger Wartell, Lively Lie, Tim Lenz, Josh Canzoneri and Arren Washington for stimulating discussions. We are grateful to Aron Yoffe for providing the program for calculating maximum ladder distances (MLDs), and to Mauricio Comas-García for bringing Reference [120] to our attention.

CHAPTER VI

ANALYSIS OF RNA SHAPE DATA

6.1 Introduction

RNA molecules are essential for life: they carry the genetic information that directs the synthesis of proteins. Some RNAs play more active roles, such as ribozymes that catalyze biochemical reactions, which require precise three-dimensional tertiary structures for proper function. According to Weeks, a "powerful way to understand these structures, especially for large RNAs in solution, is by evaluating their conformations using chemical probing technologies" [150]. Of course, there are other ways to analyze RNA structure, including X-ray crystallography, cryo-electron microscopy, and prediction. But "for many RNAs, including large structurally dynamic RNAs and conformational and functional intermediates, chemical mapping represents the best approach" [150].

One recently introduced (2005) chemical mapping technology is called selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) [93]. SHAPE is based on the premise that the nucleophilic reactivity of the ribose 2'-hydroxyl in RNA is gated by local nucleotide flexibility [151]. Thus an electrophilic reagent, such as NMIA or 1M7 [98], will react at the 2'-hydroxyl at conformationally flexible positions, forming 2'-O-adducts. The 2'-hydroxyl at nucleotides constrained by base pairing, on the other hand, will not be reactive. Sites of modification (i.e., the 2'-O-adducts) can later be determined using primer extension, followed by capillary electrophoresis and significant algorithmic analysis to generate quantitative and interpretable data [142].

In practice one cannot use SHAPE reactivity to classify whether a particular nucleotide participates in base pairing. This is because the probability distribution of SHAPE reactivity values for unpaired nucleotides, calculated from SHAPE experiments on RNAs with solved atomic-resolution structures, overlaps to a significant degree with the probability distribution of reactivity values for paired nucleotides. One can only say that there is a

clear correlation between SHAPE reactivity and Watson-Crick base pairing [15]. Despite this fundamental problem, Deigan et al. demonstrated that SHAPE reactivity information could be used to determine RNA structure with high accuracy [38]. (Though it is debatable whether SHAPE is better for secondary structure prediction than other probing technologies, e.g. DMA probing [32, 67].) This involves incorporating the SHAPE information into an energy minimization algorithm as a pseudo-free energy change term [38, 118]. SHAPE can also be used to validate structural hypotheses [66] or compare structures under different conditions [144].

The SHAPE technique relies on fast and accurate data analysis. In this chapter, we take a closer look at the data analysis pipeline. There are several publicly available programs designed to process SHAPE-CE data. These include CAFA [97], ShapeFinder [142], HiTRACE [166], FAST [111], and QuShape [65].

6.2 Overview of the SHAPE experiment

The SHAPE experimental protocol has been reported in detail elsewhere [93, 155, 153]. What follows is a brief overview of the method.

In a SHAPE experiment, we treat one sample of RNA with the SHAPE reagent. This is called the "(+) reagent" reaction, or modification reaction. We leave another sample of RNA untreated. This is called the "(-) reagent" reaction, or control or background reaction.

Next, using 5'-end-labeled DNA primers and reverse transcriptase (RT), we perform primer extension reactions on both the (+) reagent sample and the (-) reagent sample. The products of the primer extension reaction for the (+) sample will be cDNA fragments whose lengths are affected by two factors: (1) sites of modification on the RNA and (2) natural drop-off of RT. The lengths of the cDNA fragments for the (-) sample will only be affected by natural drop-off of RT.

In addition, we can carry out up to four different dideoxynucleotide (ddNTP) sequencing reactions. (The original SHAPE protocol calls for only one or two dideoxy sequencing reactions. We suppose this is due to the single-capillary approach to electrophoresis, which limits the total number of different dyes—typically, four—that can be detected in a single

capillary.) The cDNA fragments from the sequencing reactions will be used later on for nucleotide identification in our (+) and (-) samples.

Next, we use capillary electrophoresis to analyze the labeled cDNA fragments. In the original SHAPE protocol, the primers would be labeled with different color-coded fluorophores to distinguish between the modification, control, and sequencing reactions, and then the cDNAs would be combined and resolved in a single capillary on a CE instrument. The reason why the cDNAs from the modification, control, and sequencing reactions are not run in separate capillaries is because of variations in migration time from capillary to capillary, making it much more difficult to align the resulting electropherograms and leading to inaccuracies in data processing.

The final step in a SHAPE experiment is to analyze the CE electropherograms, which requires navigating a complex data processing pipeline. With the single-capillary approach to running CE, one can use the ShapeFinder program [142] to process the data. Below, we examine some alternatives to the single-capillary approach, and see how each affects the data processing pipeline.

6.3 Alternatives to the single-capillary approach

CAFA is a software program "for the high-throughput structural analysis of nucleic acids by chemical and enzymatic mapping" [97]. Although not designed specifically for SHAPE experiments, it has been used for processing SHAPE data. CAFA takes advantage of a slightly different experimental protocol: instead of running the samples in a single capillary, each sample is run in a separate capillary. Prior to running each experimental sample, a set of DNA fragments with known lengths is included with the sample. These fluorescently labeled DNA fragments are generally known as internal lane size standards, and they are subject to the same electrophoretic forces as the experimental sample (GeneScan Reference Guide). They must be labeled with a different dye so that they may be distinguished from the cDNA products in the experimental sample. Since the lengths of the size standard fragments are known, they can be used to determine the unknown lengths of the fragments in the experimental sample. So, if you can determine the lengths of the fragments, you can

assign them to the sequence. This is how CAFA works.

Like CAFA, the SHAPE data analysis program called FAST [111] uses an internal size standard so that each sample can be run in a separate capillary. The advantage of using an internal standard to normalize data among capillaries, claim Pang et al., is that the control and sequencing reactions need to be done only once, rather than repeatedly if the single-capillary approach is used. Another advantage is that a previous analysis can be used as a reference for subsequent analyses on the same RNA [65]. The FAST program also includes some new algorithms for automated y-axis scaling, signal decay correction, and peak identification.

It should be noted that this idea of using a set of fragments of known length, or size-standard fragments, to normalize variations in elution time and allow fragment identification is not new [91, 171], nor is it without problems. Wirapati has detailed many of the issues [158].

CAFA and ShapeFinder have been criticized as being inadequate for applying to large-scale titration or mutate-and-map datasets [166]. The HiTRACE program was developed to address the limitations of CAFA and ShapeFinder [166]. This required another adjustment to the experimental protocol. The idea is similar to what was done for CAFA and FAST, except instead of using a set of size standard fragments, all of the samples are co-loaded with a reference ladder derived by reverse transcribing an arbitrary RNA. Then all of the electropherograms are aligned to one another automatically using the reference ladders.

Citing the need for a program that balances processing speed, pipeline simplicity, and degree of automation, Karabiber et al. developed a new software package called QuShape [65]. One of the innovations of their approach is the use of what they call a "two-capillary protocol." What this means is that the (+) sample is co-loaded with a sequencing reaction sample in one capillary, and the (-) sample is co-loaded with the same sequencing reaction sample in another capillary. The sequencing sample is then used to do the capillary-to-capillary alignment and also to provide the sequence information.

CHAPTER VII

CONCLUSION

In our investigation of synthetic fibrin knob peptide structures and their binding with fibrin holes (Chapter 2), we combined SPR experiments and molecular dynamics simulations to determine structural properties that drive the binding dynamics. These studies provided criteria for designing knob peptides that more effectively compete for the native knob:hole interaction. The molecular dynamics simulations indicated that the 3Arg side chain orientation and peptide backbone stability were important factors in binding.

DNA has emerged as a key actor in assembling materials and devices at a very small scale. In one such application, researchers have explored using DNA to create molecular wires, with the long-term goal of producing functional nanoscale electronic devices. While DNA is not itself suitable as a nanowire due to its low conductivity, it does have properties that make it a viable candidate for nanoscale circuitry: it can be easily synthesized and, due to the specificity of Watson-Crick base pairing, it will form a duplex only if two sequences are complementary. In Chapter 3, we showed how molecular modeling techniques could be used to better understand how to design DNA-linked nanowires. We developed a computational tool that allows potential designs to be screened and evaluated—a useful tool for creating experimentally testable hypotheses. We suggested a new design based on pyrrylene vinylene monomers.

SHAPE is a powerful chemical probing technique for analyzing RNA structure. In Chapter 4, we used SHAPE to show that Domain III of the T. *thermophilus* 23S rRNA folds independently to a near-native state. This finding supports the hypothesis that Domain III was added to the ancestral ribosome as an intact entity. And in Chapter 5, we used SHAPE to study the in vitro transcript of the STMV RNA. The SHAPE-directed secondary structure we obtained was highly extended and considerably different from that proposed for the genome in the intact virion. Finally, in Chapter 6, we discussed some different

approaches to the SHAPE experimental protocol and data processing methodology that we have found helpful.

7.1 Recommendations

It seems there is never enough time to do everything you want during a scientific study. Here I address some of the questions that came up during our investigations, but were unable to be addressed in the final document, or that occurred to me after publication. These are my recommendations for future work.

From our work on fibrin knob peptides, I became very interested in how you determine if your simulation has converged. The time evolution of the root mean square deviation (RMSD) from the starting conformation of a molecular dynamics trajectory is an often-used measure of equilibrium: the system is considered to be in equilibrium when the RMSD values have stopped increasing. A low, stable RMSD trajectory reflects sampling in the vicinity of the starting conformation. Despite the popularity of using RMSD to measure equilibrium, this technique is often unable to determine when the simulation has converged. Lyman and Zuckerman recently proposed an alternative method for assessing convergence [87], and I am curious what this method would reveal about the level of sampling in our fibrin knob peptide simulations.

One of the issues that came up during the nanowire study was the issue of keeping the DNA fixed. As an improvement, one could modify the algorithm so that the DNA does not remain fixed. Alternatively, one could generate random fixed DNA conformations prior to running the loop closure algorithm, instead of using just the three conformations (A-, B-, and C-DNA) we used in our study. This would essentially involve choosing a random value for rise, twist, etc. to determine the DNA conformation.

In the nanowire study we used an algorithm for converting from torsion space to Cartesian space [112]. Many other molecular modeling algorithms require a method to convert an internal coordinate representation (bond lengths, bond angles, and dihedral angles) of a molecule into Cartesian (x, y, z) coordinates. For a protein backbone or other linear polymer chain, an obvious way to convert a list of bond lengths, bond angles, and dihedral

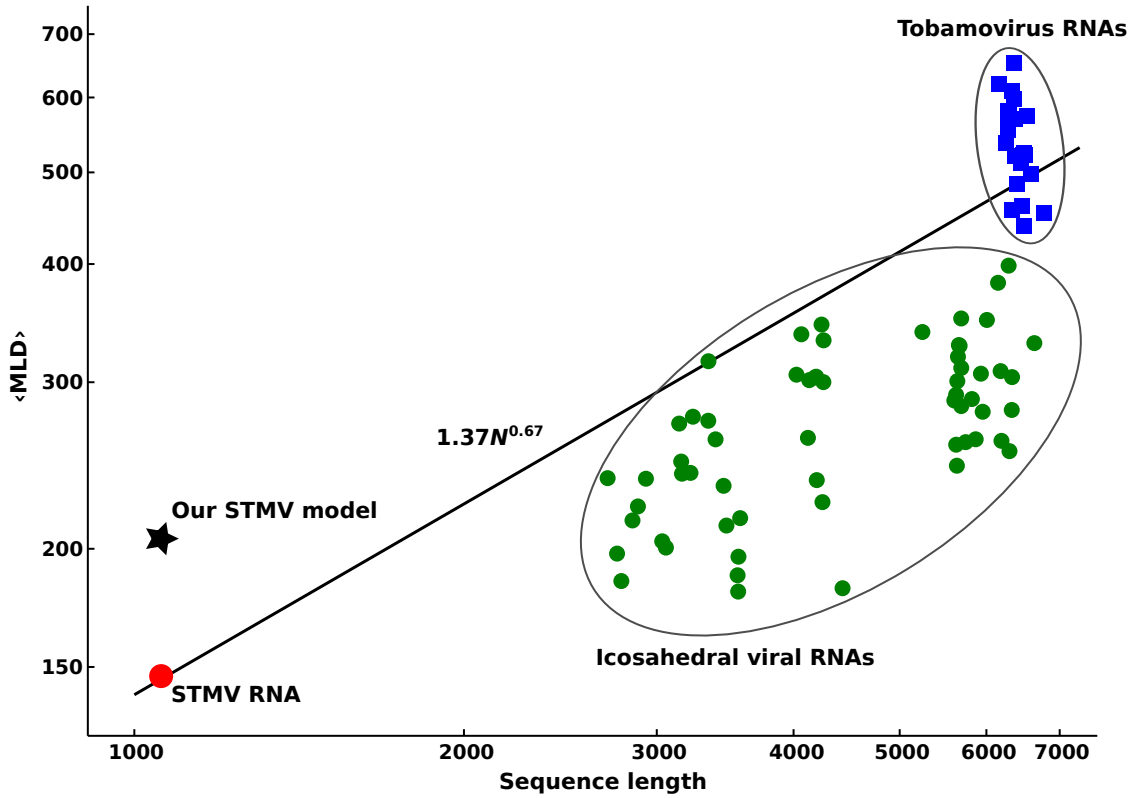


Figure 24: Log-log plot of $\langle MLD \rangle$ vs. sequence length. The green dots represent the viral RNAs of $T = 3$ icosahedral viruses analyzed by Yoffe et al. [165].

angles into Cartesian coordinates is to start at the beginning of the chain, adding one atom after another until the end is reached. Since the placement of each atom is dependent on the Cartesian coordinates of the three previously placed atoms, this computation seems inherently sequential. Using a divide-and-conquer strategy, I have developed and implemented a shared-memory parallel algorithm for performing the conversion operation. Unfortunately, the algorithm is not work-optimal, and is most effective for large problem sizes, which puts into question the scientific relevance of this idea. I have also worked on a work-optimal algorithm based on a scan operation and the Newton-Raphson algorithm for solving a system of nonlinear equations, but results so far have been discouraging. I recommend pursuing this further.

The obvious follow-up experiment to our STMV work is to perform a SHAPE experiment on STMV in the intact virus. This would afford us the opportunity to compare our results

on the in vitro transcript with results on the in virio RNA. Also, one of the more interesting findings in our STMV study was that our MLD results on STMV were at odds with the observation of Yoffe et al. that RNAs of small icosahedral viruses have smaller MLDs than do random sequences [165] (Figure 24). Is STMV simply an outlier, or does this finding generalize to other $T = 1$ viruses?

My final recommendation is in regard to SHAPE: the structural basis for the reactivity of the 2'-hydroxyl to SHAPE electrophiles is only beginning to be understood [92]. It is important that we gain a deeper understanding of the mechanism of SHAPE chemistry so that we might improve the use of SHAPE data to guide RNA secondary structure prediction [143]. I hypothesize that SHAPE reactivity will be significantly influenced by the orientation of the 2'-hydroxyl, or in other words, the C2'–O2' dihedral angle. MD simulations will allow us to examine the C2'–O2' dihedral angle, which is unknown in crystal structures since hydrogen atoms are essentially invisible in X-ray diffraction. These simulations could improve our understanding of the relationship between SHAPE reactivity and base pairing. Alternatively, using a tetraloop (i.e., a three base-pair stem with a GNRA loop) as the model system, we could perform molecular docking using the SHAPE electrophile as the ligand to study the interactions.

APPENDIX A

SUPPLEMENTAL FIGURES FOR THE FIBRIN KNOB PEPTIDE STUDY

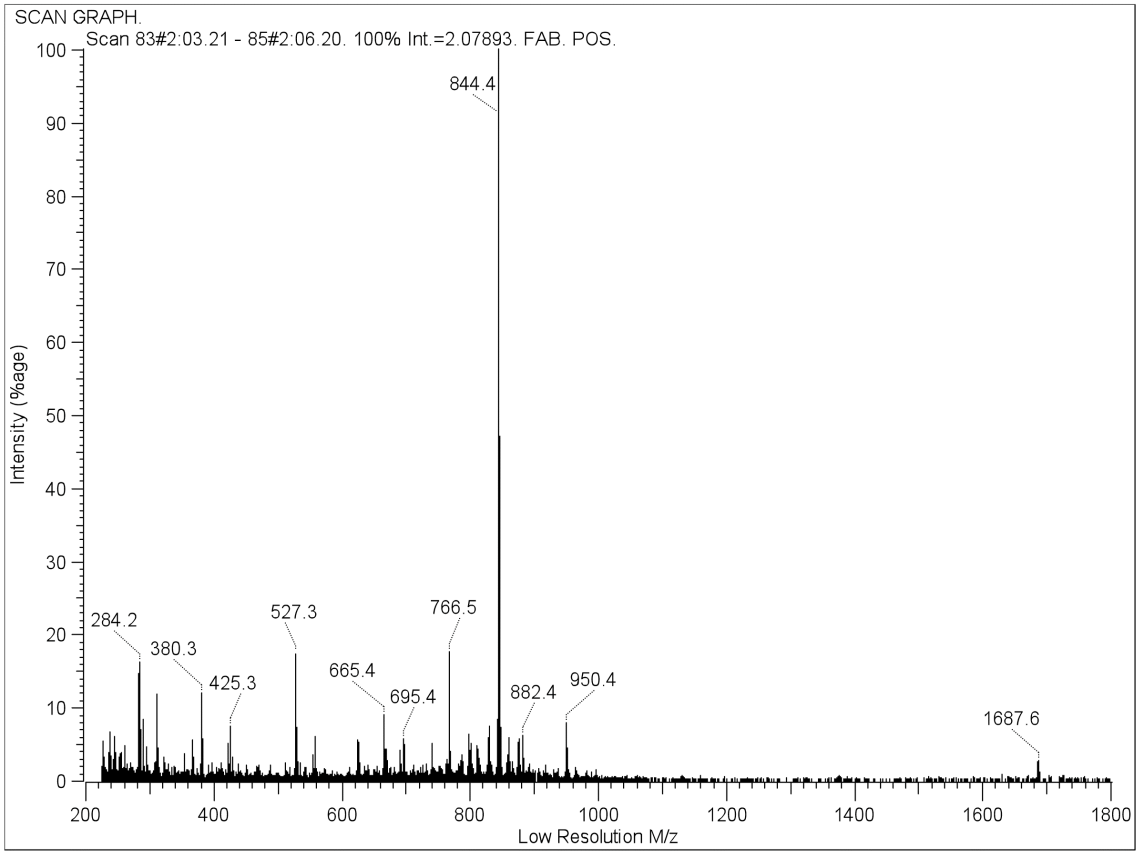


Figure 25: Representative fast atom bombardment (FAB) mass spectrometry analysis on GPRPFPAC peptide solution remaining after a SPR experiment. The analysis revealed an intense ion peak at 844.4 presumably the [M+H]+ (nominal molecular weight = 843.4 Da), indicative of a monomer peptide solution. Additionally, a minor intensity ion peak was at 1687.6, indicating minimal dimerization in the solution.

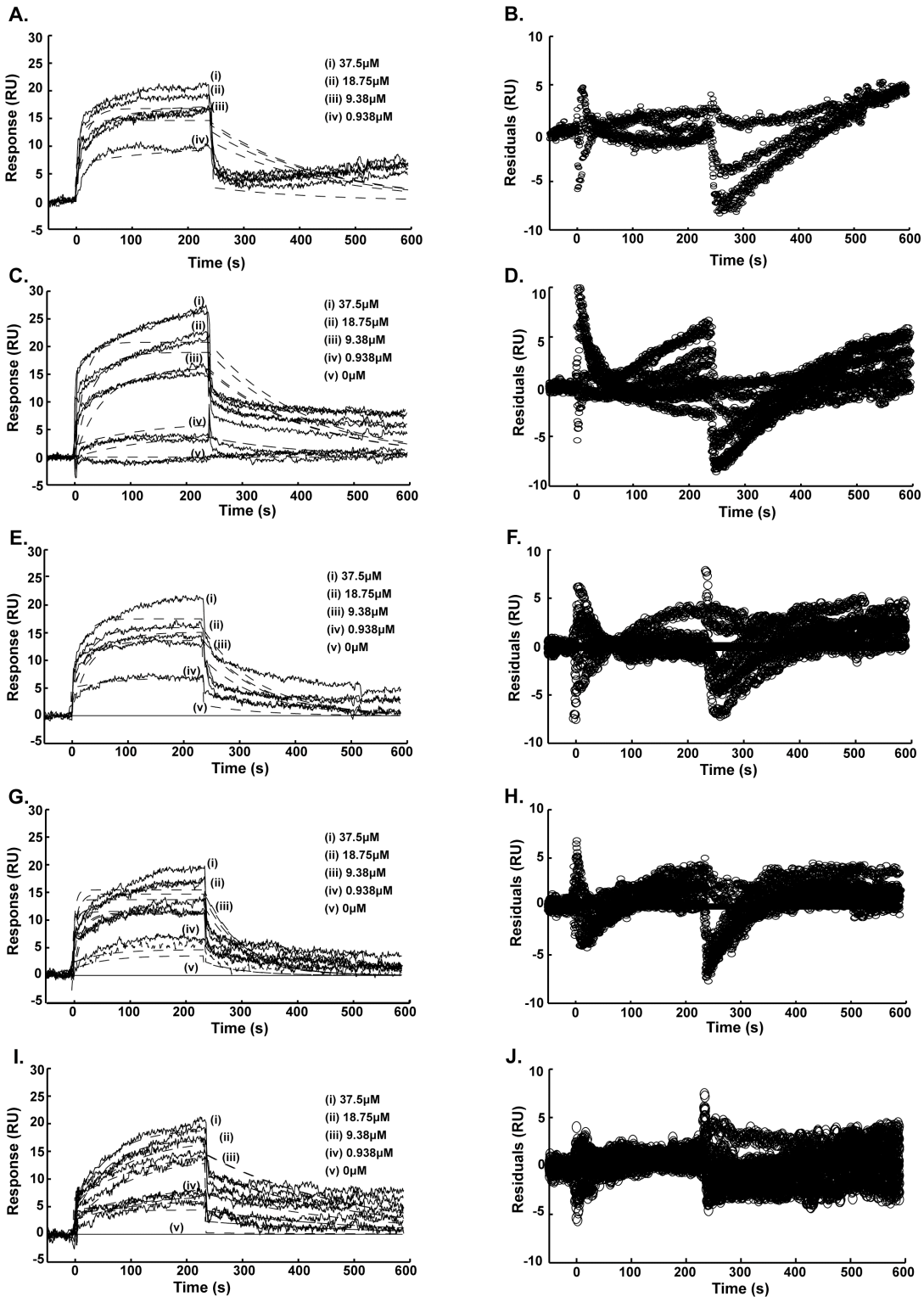


Figure 26: Experimental SPR sensorgrams with corresponding Langmuir 1:1 ligand model simulations (A, C, E, G, I) and residual plots (B, D, F, H, J); (A, B) GPRPAAC, (C, D) GPRPFPAC, (E, F) GPRPPERC, (G, H) GPRVVERC, (I, J) GPRVVAAC. Solid lines = experimental SPR response curves, dashed lines = fitted model curves.

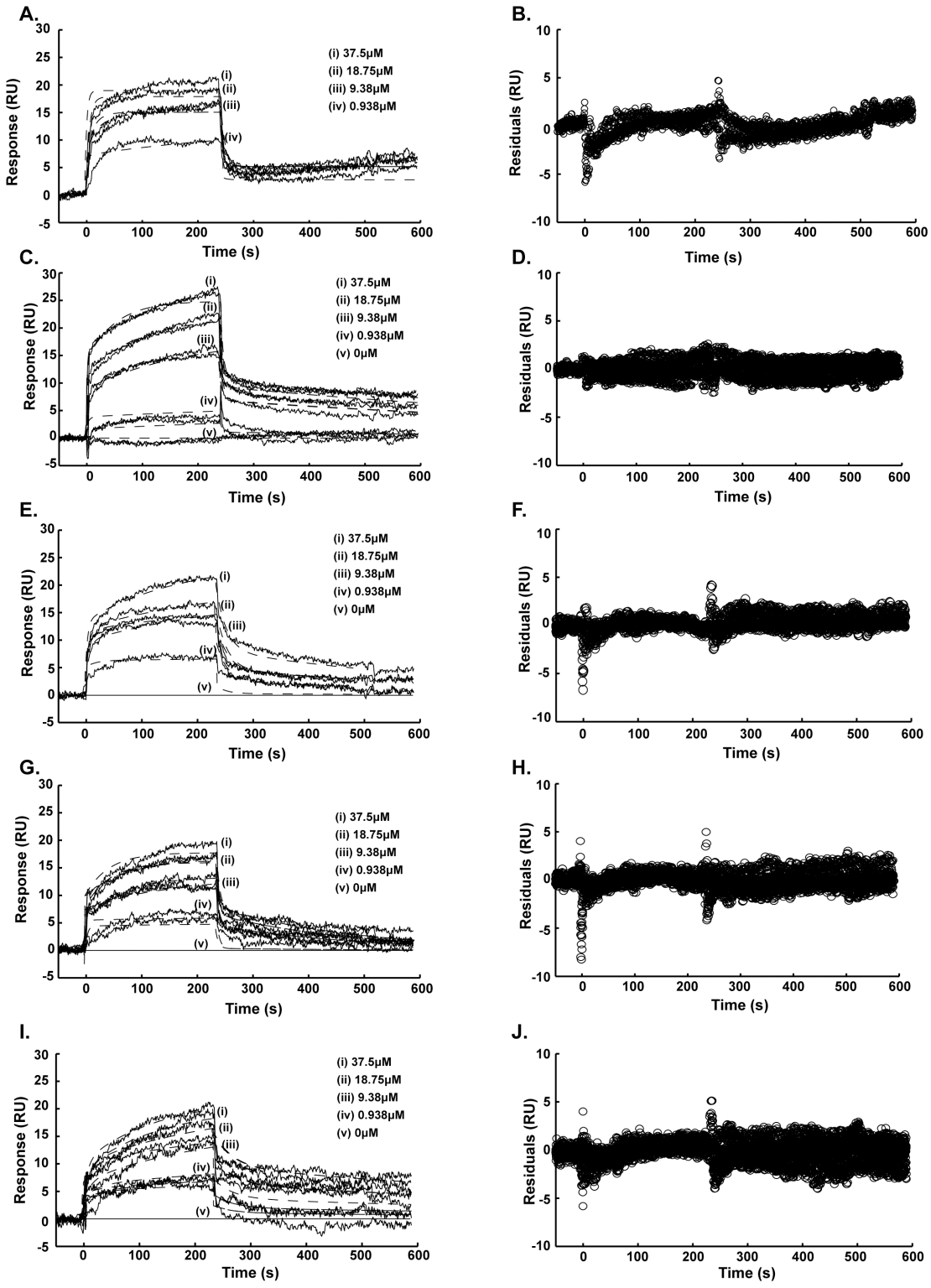


Figure 27: Experimental SPR sensorgrams with corresponding heterogeneous ligand model simulations (A, C, E, G, I) and residual plots (B, D, F, H, J); (A, B) GPRPAAC, (C, D) GPRPFPAC, (E, F) GPRPPERC, (G, H) GPRVVERC, (I, J) GPRVVAAC. Solid lines = experimental SPR response curves, dashed lines = fitted model curves.

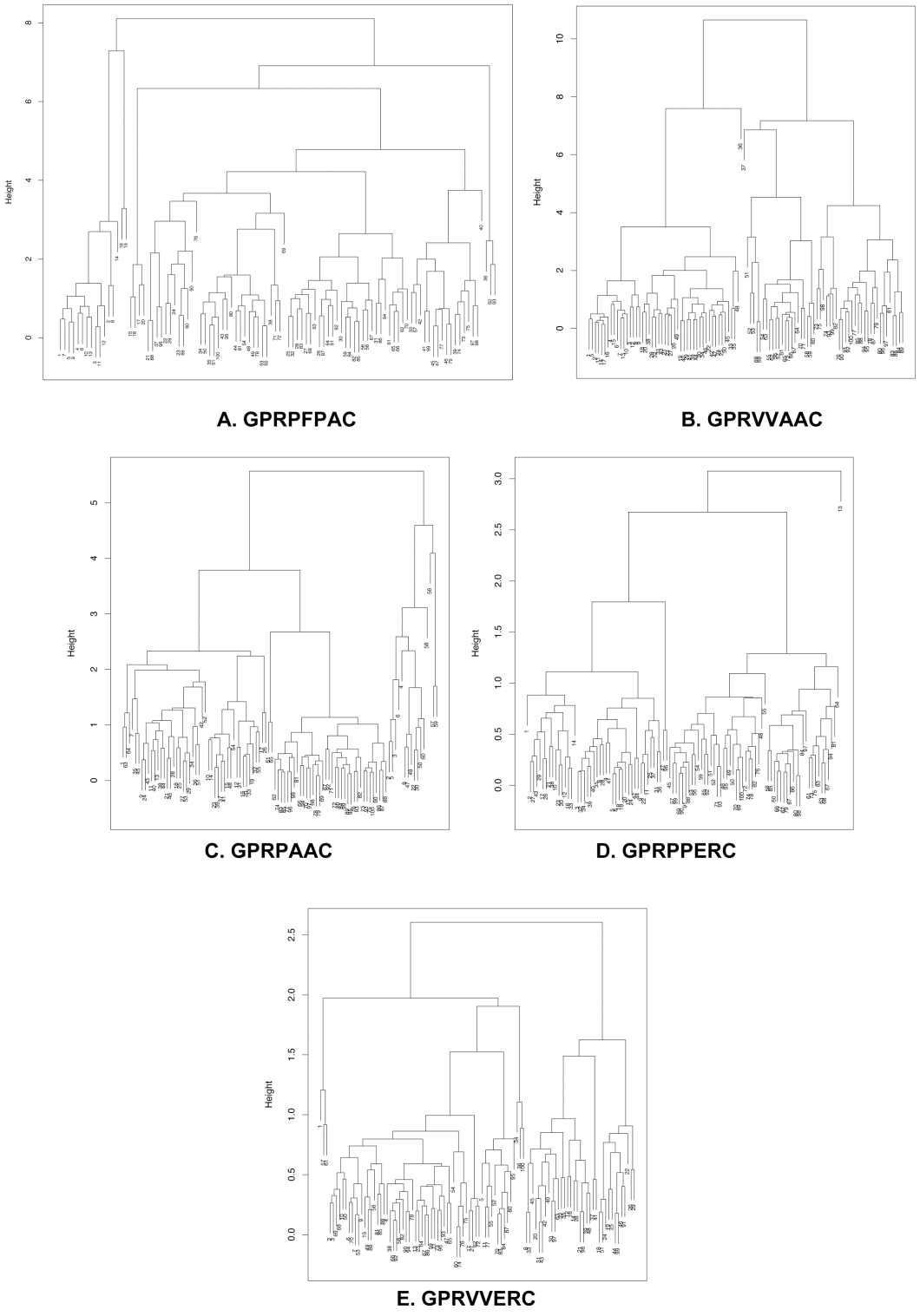


Figure 28: Dendrograms from hierarchical cluster analysis from a dissimilarity matrix generated from the RMSD between every frame in the trajectory, (A) GPRPFPAC, (B) GPRVVAAC, (C) GPRPAAC, (D) GPRPPERC, (E) GPRVVERC.

APPENDIX B

SUPPLEMENTAL MATERIAL FOR THE DOMAIN III STUDY

B.1 Results

B.1.1 SHAPE reactivity

Nineteen of the 20 most reactive nucleotides (with absolute SHAPE reactivities 54–100% of A1572, largest red triangles) are single-stranded in the canonical secondary structure of the *T. thermophilus* rRNA. The next group consisting of 20 reactive nucleotides (with absolute SHAPE reactivities 34–52% of A1572, intermediate red triangles) includes 18 nucleotides that are single-stranded in the secondary structure. Seventy-two out of 100 nucleotides in the third group of reactive nucleotides (with absolute SHAPE reactivities 5–33% of A1572, smallest red triangles) are single-stranded in the secondary structure. The fourth group consists of unreactive nucleotides of Domain III$^{\text{alone}}$ (with absolute SHAPE reactivities <2% of A1572, blue circles in Figure 10C), which is composed of 106 nucleotides out of which a large majority (eighty-two) are in double-helical regions. Nucleotides without a triangle or a blue circle in Figure 10C were not analyzed.

B.2 Materials and Methods

B.2.1 RNA Synthesis

T. thermophilus HB8 (obtained from Gaucher lab, Georgia Institute of Technology) was cultured in 0.4% yeast extract, 0.8% polypeptone peptone and 34 mM NaCl at 75°C for 96 hr with vigorous shaking. Genomic DNA was extracted using a Wizard Genomic DNA Purification kit (Promega).

Intact 23S rRNA. DNA primers complementary to the 3' ends of the *T. thermophilus* HB8 23S gene were designed with non-complementary 5' dangling ends that encode for an upstream stability sequence (5'–GTGG–3'), an EcoRI recognition site (5'–GAATTC–3'), and a T7 class II promoter f2.5 [31] (5'–TAATACGACTCACTATTAGGG–3') and a

downstream HindIII recognition site (5'–AAGCTT–3') and stability sequence (5'–GGTG–3'). Amplification of the 23S gene and addition of the dangling end sequences was completed in two steps with primers from Operon MWG. The initial PCR amplified the full 23S sequence from genomic DNA and added a portion of the total dangling end sequence. The second PCR amplified the product of the first PCR and added the remaining portion of the dangling end sequence.

The forward primer in the initial PCR was 5'–CGTAATACGACTCACTATTAGGGTC AAGATGGTAAGGGCCCAC–3' and the reverse primer was 5'–CACCAAGCTTGGAGGG GTCAAGACCTCGG–3'. The 23S was amplified from 200 ng of purified genomic DNA in 1X Fail Safe Buffer J, 2.5 U Fail Safe Enzyme (Epicentre), and 500 nM each primer. Enzyme was added 1 min into an initial denaturing step of 95°C (total, 2 min), and subsequently cycled (95°C for 1 min, 57.8°C for 1 min, 72°C for 3 min) 25 times prior to final elongation (72°C for 2 min) on an Eppendorf Mastercycler gradient thermocycler. Amplification product was purified by preparative gel electrophoresis on a 2% GTG SeaPlaque Agarose (Lonza) gel in 1X TAE at 109 V for ∼40 min, and DNA purified from gel slices by Zymoclean Gel DNA Recovery Kit (Zymo Research).

The forward primer in the second PCR was 5'–GTGGGAATTCCGTAATACGACTCAC TATTAGGGTCAAG–3' and the reverse primer was 5'–CACCAAGCTTGGAGGGGTCAA GACCTCGG–3'. This amplification was completed as described for the initial PCR except that 450 ng of gel purified amplification product from the initial PCR was the template. The product of the second PCR was purified by preparative gel electrophoresis as described above.

Approximately 1 μg of product or unmodified pUC19 vector was digested sequentially with EcoRI and HindIII (New England Biolabs): EcoRI for 1 hr at 37°C in 1X NEB EcoRI buffer followed by purification with a DNA Clean & Concentrator Kit (Zymo Research), and digestion with HindIII for 1 hr at 37°C in 1X NEB Buffer2. The digested pUC19 vector was dephosphorylated with antarctic phosphatase for 1 hr at 37°C (New England Biolabs). Enzyme was heat inactivated at 65°C (20 min). The product (digested) and vector (digested and dephosphorylated) were purified by DNA Clean & Concentrator Kit prior to 72 hr

ligation with T4 DNA Ligase (New England Biolabs) at ambient temperature. Ligation mixtures were used to transform DH5 competent E.coli cells and the resulting colonies screened after plasmid purification by secondary PCR. Colonies positive for insert were sequenced bi-directionally for consensus (Operon MWG). Point mutations were corrected by site-directed mutagenesis (Agilent), and corrections confirmed by further sequencing. The completed construct is subsequently referred to as the 23S/pUC19 construct.

Domain III rRNA. Domain III is defined here as RNA residues G1271–G1647 of the *T. thermophilus* 23S RNA secondary structure numbered with the E.coli numbering scheme [23, 167]. Primers complementary to the 23S DNA template strand equivalent of RNA residues G1271–G1296, and non-template strand residues G1626–G1647 were designed without and with non-complementary 5' dangling ends. Dangling ends were identical to those designed for the intact 23S except that a standard T7 promoter (5'–TAATACGACTCACTATAGGG–3') was used. The Domain III gene was amplified and dangling ends appended in a two-step process.

The forward primer in the initial PCR was 5'–GATAAAGAGGGTGAGAATCCCTCT CG–3' and the reverse primer was 5'–CGCGCCTGAGTGCTCTTGCACC–3'. Domain III was amplified from 100 ng of 23S/pUC19 construct in 1X Cloned Pfu DNA Polymerase Buffer (Agilent Technologies), 250 μM each dNTP (New England Biolabs), 500 nM each primer and approximately 3 U PFU DNA polymerase (Agilent Technologies). The PCR was cycled as described for the 23S but with an annealing temperature of 55°C and extension time of 25 s. The amplification product was purified by DNA Clean & Concentrator Kit.

In the second PCR, the forward primer was 5'– GTGGGAATTCTAATACGACTCAC-TATAGGGATAAAGAGGGTGAGAATCCCTC TCG–3' and the reverse primer was 5'–CACCAAGCTTCGCGCCTGAGTGCTCTTGCACC–3'. This reaction was assembled and cycled as described for the first Domain III PCR except that 100 ng of purified product from the first PCR was used as template. PCR product was recovered from a 2% preparative agarose gel and ligated into pUC19 as described above. Ligation mixtures were used to transform DH5α competent *E.coli* cells and the DNA sequenced as described above.

Transcription. Transcription reactions were performed by the run-off method [124],

using the MEGAscript High Yield Transcription Kit (Applied Biosystems). pUC19 constructs containing either the complete 23S sequence or the 23S Domain III sequence were linearized by digestion with HindIII and purified by DNA Clean & Concentrator Kit as described above. Linearized pUC19/23S construct (0.5 μg) was transcribed in 20 μL reaction volumes for 4 hours at 37°C. Linearized pUC19/Domain III construct (1 μg) was transcribed in 20 μL reaction volumes for 16 hours at 37°C. Transcription reaction conditions were scaled as appropriate to optimize purity and yield. RNA products from transcription reactions were recovered by ammonium acetate precipitation and resuspended in nuclease-free water (IDT). Yields were quantified by UV absorbance.

B.2.2 SHAPE data processing

The output of a SHAPE experiment is a series of capillary electrophoresis (CE) data traces, or electropherograms, which report fluorescence intensity values as a function of migration time [142]. CE traces were converted into final SHAPE reactivity measurements with in-house Matlab code, adapted from a previous protocol [142, 153]. Seven key steps were required: (i) alignment, (ii) baseline correction, (iii) sequence assignment, (iv) peak quantification, (v) signal decay correction, (vi) background subtraction, and (vii) normalization.

Peak alignment is essential for extracting meaningful reactivity measurements, as well as for achieving reproducibility between experiments. Traces were aligned to one another with the help of fluorescently labeled internal-lane size standards (DNA ladders) that were co-loaded with the reverse transcription (RT) reaction products. One DNA ladder trace was used as the reference to which the other ladder traces were aligned. Peak locations in the traces were found automatically, and then manually adjusted in some cases. The key was to locate matching peaks in all traces.

A modified version of a technique known as parametric time warping [45] was used to transform the time axis so that peaks in two traces would be aligned. Consider the alignment of a given trace $y(x)$ to the reference trace $y_{ref}(x)$. Let u be the vector of peak locations in y, and let u_{ref} be the vector of peak locations in y_{ref}. The difference $d = u - u_{ref}$ between these vectors represents the offset of the peaks in y relative to the peaks in y_{ref}. This offset

was fit with a polynomial $p(x)$ of degree 3 such that the difference between $p(u(i))$ and $d(i)$ was minimized in the least squares sense. The aligned ladder trace y' was then computed by nearest-neighbor interpolation of the intensity values in y given the transformed x-values, i.e., $x - p(x)$. Since the DNA ladder was co-loaded with the RT reaction products, the same correction was applied to the corresponding RT reaction trace.

Electropherograms typically contain a slowly varying component that imparts a vertical offset to the baseline, that, if not corrected for, can lead to inaccurate peak quantification [2, 142]. The following procedure, originally described by Berno [12], was used to adjust the baseline of each trace. The baseline was estimated by first computing the fifth percentile of signal intensity values in overlapping windows of the data, and then constructing a piecewise linear signal by linear interpolation of the percentile values found in successive windows. The window size was set to approximately 20 times the average peak width, and the window spacing was set to half the window size. The final baseline was obtained by smoothing the piecewise baseline with a Gaussian filter. This baseline was subsequently subtracted from the original trace to get the baseline corrected data.

The dideoxy sequencing traces are necessary for mapping the peaks in the SHAPE reaction traces to the RNA sequence [142]. First, peaks in the the dideoxy sequencing traces were located automatically and then assigned to the sequence. Next, peaks were located in the SHAPE reaction traces and linked to peaks in the sequencing traces. Peak assignments were monitored by visual inspection and corrected as required. Since the cDNAs in the sequencing reactions are one nucleotide longer than the corresponding cDNAs in the SHAPE reactions, the final reactivity measurements were shifted by one nucleotide [142].

The RT reaction products were quantified by computing the area of each peak. As shown previously [142], a peak can be accurately modeled by a Gaussian function, and furthermore an entire trace $y(x)$ can be modeled as a sum of K Gaussians [166], one curve for each peak: $f(x) = \sum_{k=1}^{K} a_k \exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right)$ where a_k is the amplitude, μ_k is the center, and σ_k is the width of a particular Gaussian. The optimal values of these parameters were found by least squares minimization. To speed up the calculations, a sliding window approach was used.

The window size was set to contain 15 peaks ($K = 15$), and the windows were spaced every 5 peaks. The final peak area was given by $A_k = a_k \sigma_k \sqrt{2\pi}$.

Due to (a) imperfect processivity of reverse transcriptase and (b) some RNA molecules containing more than one 2'-*O*-adduct, short cDNAs are overpopulated after the RT reaction, leading to a falloff in the observed intensity values as the read length increases [7, 142]. This so-called signal decay, which is of course reflected in the peak areas, was modeled with an exponential function [7]: $f(i) = aq^i + b$ where a is the amplitude of the falloff, q is the probability of extension, b is the intensity offset, and i is the nucleotide position. These parameters were found by nonlinear least-squares curve-fitting. Outliers, defined here as the top 2% of the peak areas, were excluded from the curve-fitting procedure. The corrected peak areas were then calculated as $A'(i) = A(i)/f(i)$.

To account for small differences in signal intensity between different runs of the CE instrument, the traces from the SHAPE reagent reactions were scaled such that the smallest 40–50% of the peaks matched the corresponding no-reagent (background) reaction peaks [142]. The optimal scale factor was found by minimizing the sum of the absolute differences between the scaled areas and the background areas. Finally, the background was subtracted from the scaled areas to obtain the SHAPE reactivity values.

As a final step, the SHAPE reactivity values were normalized to a uniform scale according to the following procedure by Low and Weeks [86]. Briefly, after ignoring the most reactive 2% of all reactivity measurements, the next most reactive 8% of the measurements were averaged, and then all measurements were divided by this average.

B.2.3 Tertiary interactions

The program RNAview [164] was used to identify the numbers and types of inter- and intra-domain interactions for Domain III in the *T. thermophilus* 23S rRNA. The program determines types of interactions in an RNA structure using various standard geometrical references. Each interaction is classified according to Leontis and Westhof's definition [78] along with the Saenger nomenclature [122]. Leontis and Westhof annotated RNA motifs into 12 geometric families based on their distinct edge-to-edge interactions (Watson-Crick,

Hoogsteen or Sugar edge) along with the orientation of glycosidic bond (cis or trans) [79]. Saenger legend describes 28 possible base pairs between A, G, C, or U (T), which involve at least two H-bonds [122]. To avoid any misinterpretations, tertiary interactions were redefined as any types of interactions that could not be inferred directly from the secondary structure. Tertiary interactions include long-range base-base stacking and long-range base-base hydrogen bonds.

Long-range base-base H-bonds identified by RNAview were checked to see if they involve at least one H-bond as defined by Leontis and Westhof [78]. For this purpose the "FindHBond" plugin in Chimera was used [94, 114]. H-bond criteria used was D...H distance $<$ 4.0 Å, and X–D...A angle $<30°$. Note that RNAView uses only the distance criterion to find H-bonds. The classification of the interaction edges and the presence of at least one H-bond in these tertiary interactions were confirmed from the crystal structure of the *T. thermophilus* 50S ribosomal subunit. The stacking interactions were also verified by inspection of the structure to ensure the correct interaction classification. Tables 5 and 6 report intra- domain tertiary interactions for Domain III. Tables 7 and 8 report the inter-domain interactions for Domain III. Watson Crick pairs are denoted as $-/-$ (AU) or $+/+$ (GC). Other types of interactions are reported according to their interacting edges: W(Watson-Crick), H(Hoogsteen), and S(sugar). In Tables 7 and 8, the first column (Resid_i) represents the RNA residues that belong to Domain III. Tables 5–8 also show whether SHAPE detects these interactions as indicated by calculating the difference between SHAPE reactivity with and without Mg^{2+}. A difference of 15% or higher was considered as an indication that a tertiary interaction is detected by SHAPE.

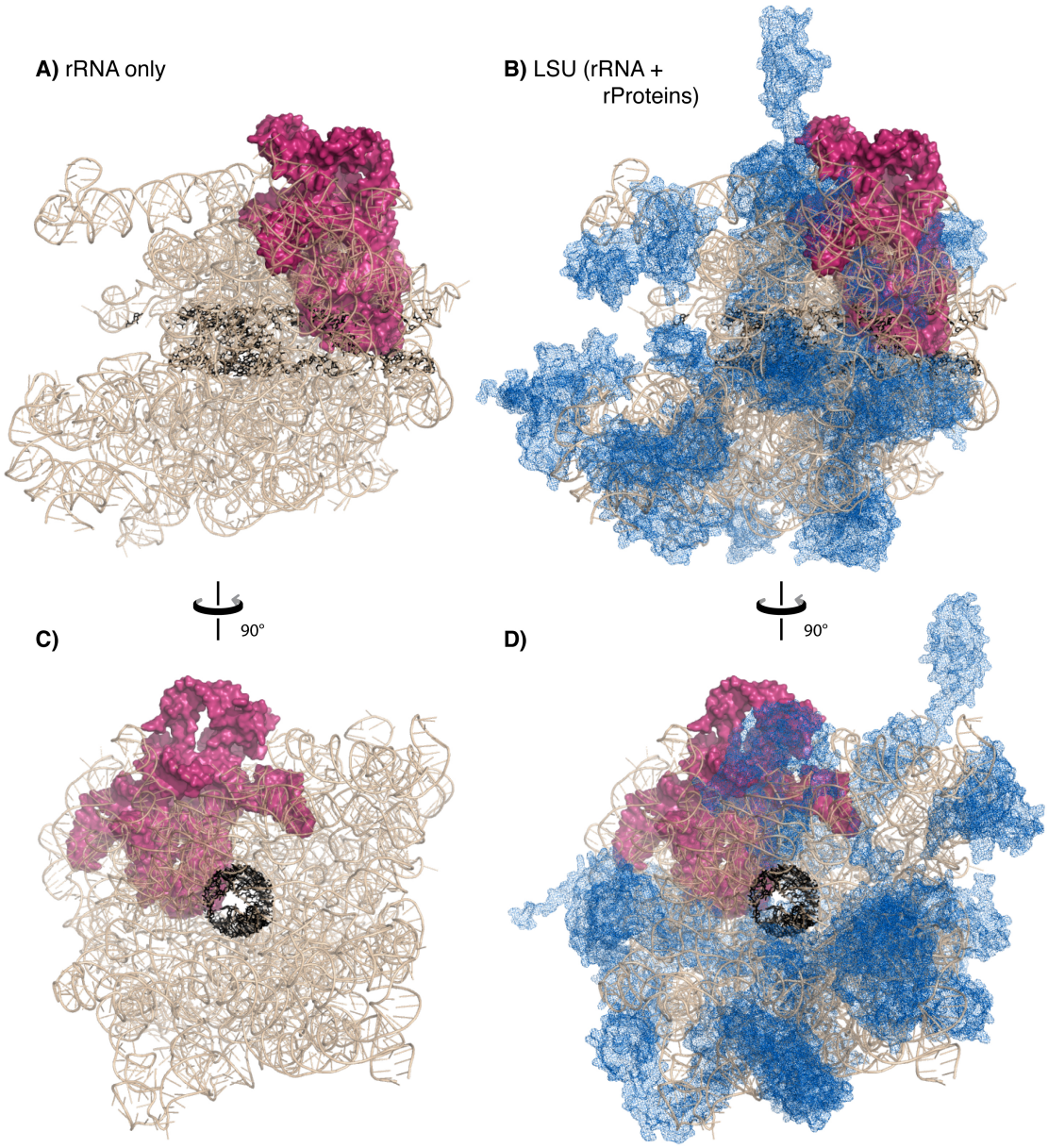


Figure 29: *T. thermophilus* LSU, with Domain III highlighted in a pink surface representation. To assist in establishing orientation, the rRNA atoms lining the peptide exit tunnel are highlighted in black. Other rRNA (light brown) is represented by semitransparent cartoon. A) rRNA only viewing across the peptide exit tunnel. B) rRNA + rProteins (blue mesh). C) rRNA only viewing down the peptide exit tunnel. D) rRNA + rProteins.

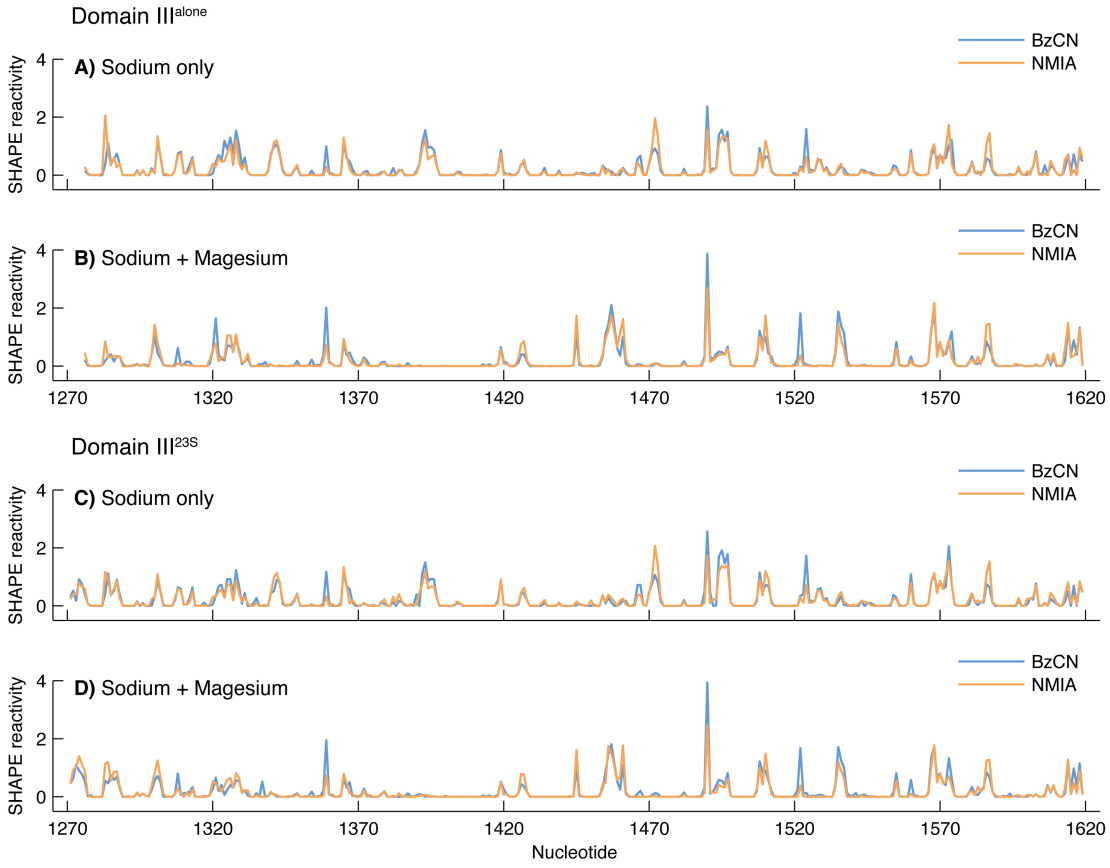


Figure 30: SHAPE reactivities for Domain III$^{\text{alone}}$ (panels A and B) and Domain III^{23S} (panels C and D) obtained using N-methylisatoic anhydride (NMIA) and benzoyl cyanide (BzCN). The vertical axis represents SHAPE reactivities and the horizontal axis represents nucleotide position.

Table 5: Intra-domain tertiary interactions of Domain III. Watson Crick pairs are denoted as -/- (AU) or +/+ (GC). Other types of interactions are reported according to their interacting edges: W(Watson-Crick), H(Hoogsteen), and S(sugar). The right most column specifies whether SHAPE identifies these interactions: ✓ indicates yes, ✖ indicates no, and NA indicates that SHAPE reactivity was unavailable.

#	Resid_i (2J01)	Resid_j (2J01)	i-j bases	Interaction edges	Orientation	Saenger Legend	# of H-bonds	SHAPE
1	1271	1615	G-C	W/W	tran	XXII	2	1271 NA, 1615 ✖
2	1274	1296	A-G	S/S	cis	n/a	1	1274 NA, 1296 ✖
3	1274	1644	A-C	W/S	tran	!(b_s)	1	1274 NA, 1644 NA
4	1275	1645	A-G	S/S	cis	n/a	3	1275 ✓, 1645 NA
5	1281	1286	G-A	W/H	tran	n/a	1	1281 ✖, 1286 ✖
6	1283	1329	G-U	S/W	cis	n/a	2	1283 ✓, 1329 ✓
7	1288	1326	U-U	W/W	tran	XII,XIII	2	1288 ✖, 1326 ✖
8	1297	1302	C-A	S/W	tran	!(b_s)	1	1297 ✖, 1302 ✓
9	1299	1640	G-C	S/H	tran	!(s_s)	1	1299 ✖, 1640 NA
10	1301	1626	A-G	S/S	tran syn	n/a	1	1301 ✓, 1626 NA
11	1301	1641	A-A	H/W	tran syn	V	2	1301 ✓, 1641 NA
12	1308	1608	A-A	W/W	tran syn	I	2	1308 ✓, 1608 ✓
13	1309	1611	G-C	S/S	cis	!(s_s)	1	1309 ✓, 1611 ✖
14	1310	1610	G-A	S/S	cis	n/a	1	1310 ✖, 1610 ✖
15	1312	1340	U-U	W/W	tran	!1H(b_b)	1	1312 ✖, 1340 ✖
16	1315	1392	C-A	S/S	cis	!(s_s)	2	1315 ✖, 1392 ✓
17	1320	1331	C-A	W/H	tran	XXV	2	1320 ✓, 1331 ✓
18	1322	1333	A-C	W/S	cis	n/a	2	1322 ✖, 1333 ✖
19	1327	1647	C-G	W/S	tran syn	n/a	2	1327 ✓, 1647 NA
20	1338	1393	G-A	S/S	tran	n/a	2	1338 ✖, 1393 ✓
21	1343	1384	G-A	S/S	tran	n/a	2	1343 ✓, 1384 ✖
22	1343	1404	G-C	+/+	cis	XIX	3	1343 ✓, 1404 ✖
23	1344	1403	G-C	+/+	cis	XIX	3	1344 ✖, 1403 ✖
24	1384	1404	A-C	S/S	cis	!(s_s)	2	1384 ✖, 1404 ✖
25	1388	1525	G-G	S/S	cis	!(s_s)	1	1388 ✖, 1525 ✖
26	1391	1393	U-A	S/S	tran	!(s_s)	1	1391 ✓, 1392 ✓
27	1417	1587	C-A	S/S	cis	!(s_s)	1	1417 ✖, 1587 ✖
28	1422	1492	G-G	S/S	cis	!(s_s)	1	1422 ✖, 1492 ✖
29	1422	1498	G-C	S/S	tran	!1H(b_b)	1	1422 ✖, 1498 ✖
30	1423	1499	G-C	S/S	cis	!(s_s)	1	1423 ✖, 1499 ✖
31	1437	1516	C-C	S/S	cis	!(s_s)	1	1437 ✖, 1516 ✖
32	1448	1528	G-A	S/H	tran syn	!(b_s)	1	1448 ✖, 1528 ✓
33	1448	1528A	G-A	S/W	tran	X	1	1448 ✖, 1528 ✓
34	1449	1463	A-C	W/S	cis	!1H(b_b)	1	1449 ✖, 1463 ✖
35	1449	1530	A-C	S/H	tran	!(b_s)	1	1449 ✖, 1530 ✓
36	1465	1545	G-A	S/W	tran	X	2	1465 ✓, 1545 ✖
37	1478	1558	G-A	S/W	tran syn	!(s_s)	1	1478 ✖, 1558 ✖
38	1514	1557	U-C	S/S	cis	!(b_s)	1	1514 ✖, 1557 ✖
39	1515	1556	G-C	S/S	cis	!(s_s)	1	1515 ✖, 1556 ✖
40	1554	1634	A-A	W/H	tran	V	2	1554 ✓, 1634 ✖
41	1604	1610	C-A	S/W	tran	!(b_s)	1	1604 ✖, 1610 ✖
42	1607	1622	C-G	W/H	tran	n/a	1	1607 ✖, 1622 ✖

Table 6: Intra-domain stacking interactions of Domain III.

#	Resid_i (2J01)	Resid_j (2J01)	i-j bases	SHAPE
1	1276	1295	A-C	1276 ✖, 1295 ✖
2	1277	1294	G-U	1277 ✖, 1294 ✖
3	1278	1293	A-C	1278 ✖, 1293 ✖
4	1288	1327	U-C	1288 ✖, 1327 ✓
5	1300	1626	U-G	1300 ✓, 1626 NA
6	1300	1634	U-A	1300 ✓, 1634 NA
7	1313	1610	U-A	1313 ✓, 1610 ✖
8	1328	1330	G-C	1328 ✓, 1330 ✓
9	1343	1597	G-A	1343 ✓, 1597 ✓
10	1349	1598	A-C	1349 ✓, 1598 ✖
11	1350	1382	C-G	1350 ✖, 1382 ✖
12	1442	1550	G-C	1442 ✖, 1550 ✖
13	1487	1503	G-U	1487 ✖, 1503 ✖
14	1492	1499	G-C	1492 ✖, 1499 ✖

Table 7: Inter-domain tertiary interactions of Domain III.

#	Resid_i (2J01) Domain III	Resid_ j (2J01)	i-j bases	Interaction edges	Orientation	Saenger Legend	# of H-bonds	Interacting domain	SHAPE
1	1365	187	A-G	S/S	cis	!(s_b)	1	Domain I	✓
2	1366	210	A-C	S/s	cis	!(s _s)	1	Domain I	✗
3	1407	142	C-A	S/s	cis syn	!(s _s)	2	Domain I	✗
4	1408	141	C-A	s/S	cis syn	!(s_b)	1	Domain I	✗
5	1595	142	G-A	S/H	tran syn	XI	1	Domain I	✗
6	1353	693	A-C	S/s	cis	!(s _s)	2	Domain II	✗
7	1354	692	A-C	s/S	cis	!(s _s)	1	Domain II	✗
8	1354	770	A-G	S/S	tran	n/a	1	Domain II	✗
9	1355	771	G-G	S/S	cis	!(s _s)	1	Domain II	✗
10	1632	700	A-G	S/s	cis	!(s _s)	2	Domain II	NA
11	1633	699	G-A	s/S	cis	!(s _s)	1	Domain II	NA
12	1287	1648	A-C	W/S	cis	n/a	2	Domain IV	✓
13	1288	1647	U-G	S/W	tran syn	!(s _s)	1	Domain IV	✗
14	1326	2010	U-G	S/S	cis	!(s _s)	1	Domain IV	✓
15	1327	1647	C-G	W/S	tran syn	n/a	2	Domain IV	✓
16	1362	1810	C-A	S/s	cis	!(s _s)	1	Domain IV	✗
17	1369	1809	G-A	S/W	tran	!(b _s)	1	Domain IV	✗
18	1369	1810	G-A	S/S	tran	n/a	1	Domain IV	✗
19	1550	1743	C-C	S/S	cis	!(s _s)	1	Domain IV	✗
20	1631A	1682	A-G	H/S	tran	!(b _s)	1	Domain IV	NA
21	1635	1761	G-C	S/S	cis	!1H(b _b)	1	Domain IV	NA
22	1636	1760	C-A	s/S	cis	!(s _s)	2	Domain IV	NA
23	1455	2704	G-C	+/+	cis	XIX	3	Domain VI	✓
24	1456	2703	G-C	+/+	cis	XIX	3	Domain VI	✗
25	1638	2698	C-U	s/S	tran	!(s _s)	1	Domain VI	NA

Table 8: Inter-domain stacking interactions of Domain III.

#	Resid_i (2J01) Domain III	Resid_j (2J01)	i-j bases	Interacting domain	SHAPE
1	1456	2704	G-C	Domain VI	✗
2	1457	2703	A-C	Domain VI	✗

APPENDIX C

SUPPORTING INFORMATION FOR THE STMV STUDY

C.1 Methods S1

We used 5 DNA primers to analyze the STMV RNA. They were designed to anneal at roughly equally spaced positions along the sequence so that the combined primer extension reactions would span the entire RNA (Table 9). The primers were designed with the assistance of Primer3Plus [140]. Since we typically obtain read lengths of more than 300 nucleotides per primer extension reaction, the primer extension reactions for STMV RNA resulted in regions of overlapping data from different primers. These regions of overlapping data are important for our data processing procedure. Note that for this part of the procedure we number the nucleotides from 1 to 1058 with respect to the 3' end, not the 5' end.

When using multiple primers to analyze an RNA, the typical approach is to process the data from each primer extension reaction individually [38, 149]. We take a similar approach here to convert the capillary electrophoresis data into raw peak areas. But after this step we deviate from the established protocol and combine the peak area data from all the primer extension reactions into one signal. We find it easier to complete the processing steps of correcting for signal decay, subtracting the background, and normalizing the data if we are working with one combined dataset.

We combine the data by taking advantage of the information contained in the regions of overlapping data. Plotting the peak area signals for all of the individual primer extension datasets on a single plot, we see that the data in the overlapping regions do not match up (Figure 31, top panel). In other words, the data from one primer extension reaction will be higher or lower than the one that it overlaps with. There are two reasons for this. First, the data from two different primer extension reactions will not in general be on the same scale due to experimental variations. Second, the signal for each primer extension reaction decays

Table 9: Primers used to analyze the STMV RNA.

Primer	Primer annealing location	Nucleotides read (useable data)	Nucleotides used in combined signal
1	1–20	25–370	25–236
2	209–228	237–595	237–396
3	372–391	397–756	397–655
4	629–648	656–1014	656–864
5	839–858	865–1053	865–1053

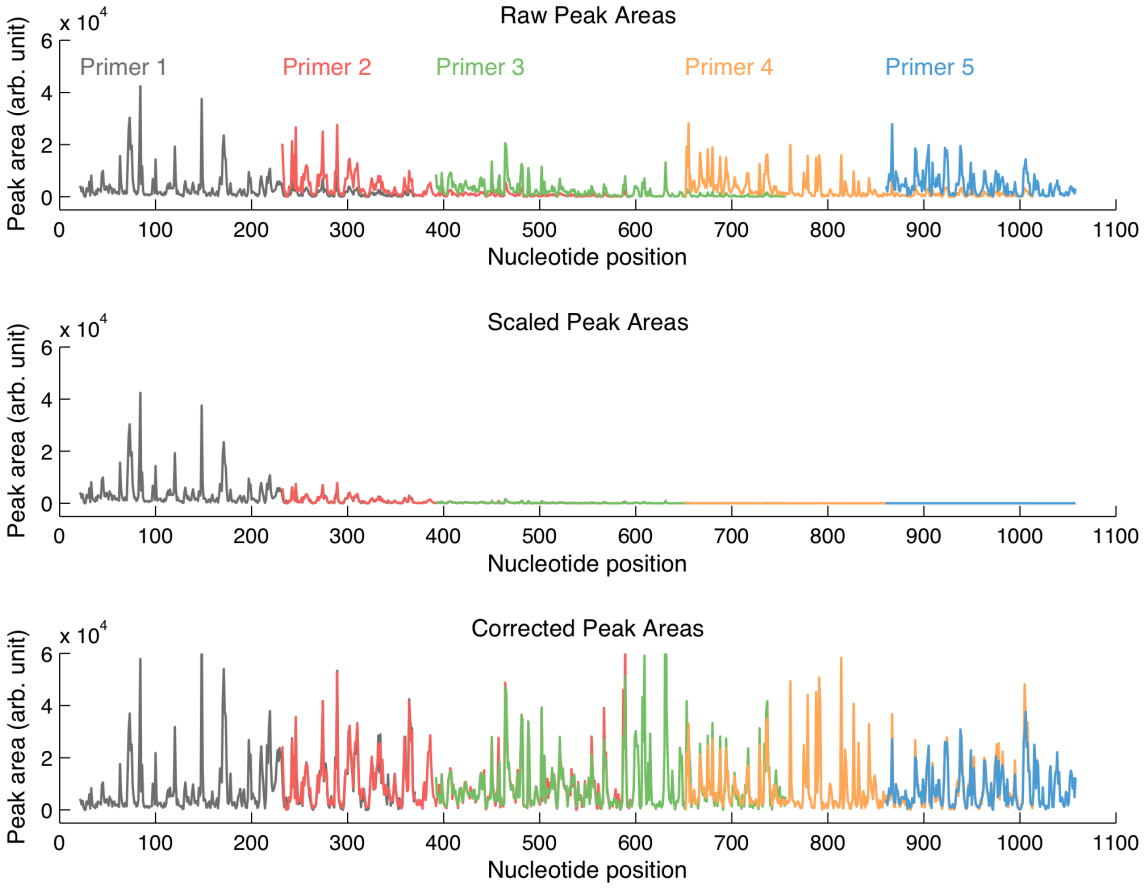


Figure 31: Signal decay correction. The regions of overlapping data from different primers are not on the same scale (top). After scaling all of the primers to one another such that the overlapping regions match up, the resulting signal decays rapidly (middle). After correcting for signal decay, the overlapping regions are in agreement (bottom).

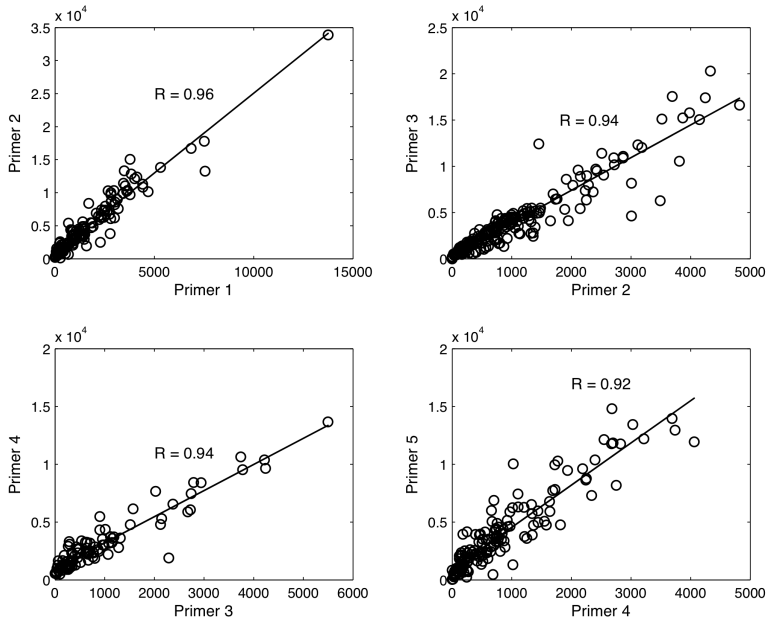


Figure 32: Quantitative correlation between peak area data in overlapping primer reads. This demonstrates that signal decay in the regions of overlapping data is similar. Pearson's r-values are shown.

in an approximately exponential fashion for reasons that have been explained previously [6, 142]. We observe here that whatever factors cause signal decay in one primer extension reaction should also apply to the other primer in the region of overlap. This is confirmed by computing Pearson's correlation coefficient for the peak areas in the overlapping region between two different primer extension reads (Figure 32). As expected, we see a linear relationship. Therefore, we need only to apply a scaling factor to one of the primer datasets to have the overlapping regions match up. We do this by automatically finding the scaling factor that minimizes the sum of squares difference between the peak areas in one primer dataset and the corresponding peak areas in the overlapping primer dataset. For example, we scale the primer 2 data to the primer 1 data so that the peak areas match. Then we scale the primer 3 data to the scaled primer 2 data, and so on until we have scaled all the primer data (Figure 31, middle panel). To combine the data from all of the primers into one signal, we use data from each of the primers as shown in Table 9. We use primer 1 data up until the point primer 2 starts, and then we use primer 2 data up until the point primer 3 starts, and so on. There are other ways of combining the data, for example by taking the

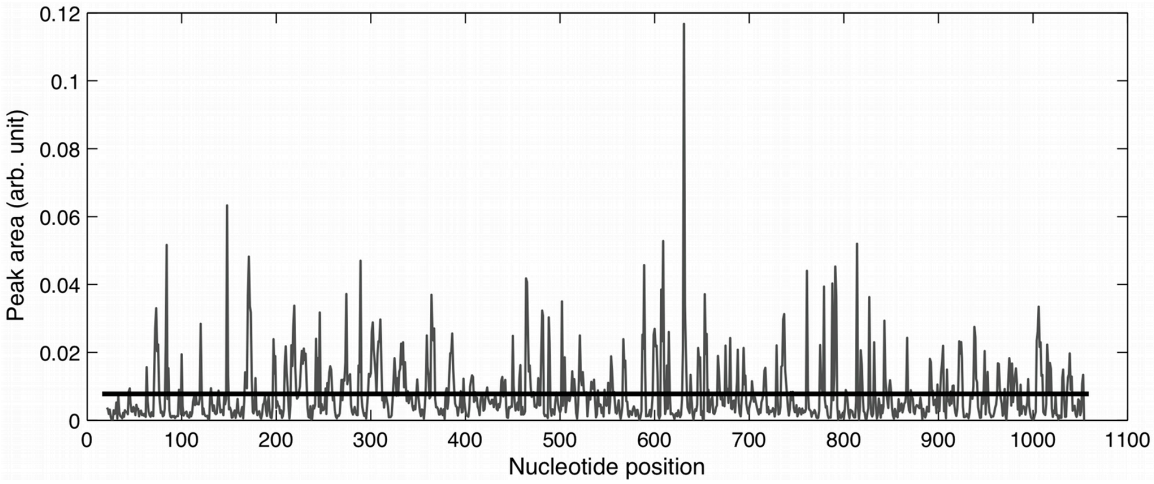


Figure 33: Combined peak area signal after decay correction. The thick black line fitted to the corrected peak area data has a slope of zero, ensuring that intense values in the beginning, middle, and end of the signal are of uniform height.

average of the values in the overlapping region. The combined dataset spans nucleotides 25 to 1053.

Next we perform signal decay correction on the combined dataset. Rather than fitting the data to an exponential function, we use a nonparametric correction factor developed by Aviran et al. for modeling polymerase drop-off [6]. The corrected peak area, Y_k, is calculated as follows:

$$Y_k = \log\left(1 - \frac{X_k}{\sum_{i=k}^{n+1} X_i}\right) \tag{4}$$

where n is the number of nucleotides and X_k is the raw peak area for nucleotide k for $k = 1, \ldots, n$. The amount of full-length transcript is represented by X_{n+1}. Since we are not generally able to quantify the amount of full-length transcript, we must approximate its value. We do this by fitting a straight line to the corrected data and choosing the value for X_{n+1} that results in a line with a slope of zero (Figure 33). The intense values in the beginning, middle, and end of the signal are thus of uniform height [142].

The remaining data processing steps are performed as described previously [4].

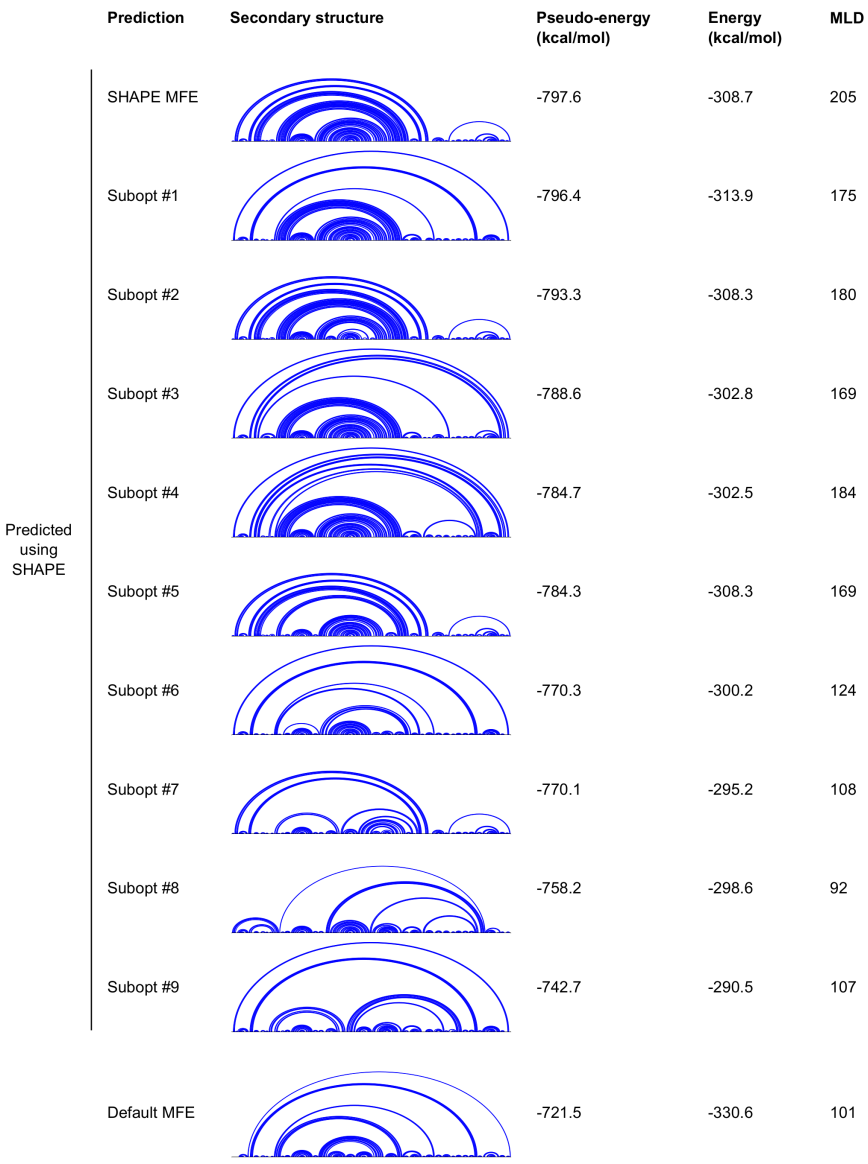


Figure 34: Predicted secondary structures for STMV RNA. SHAPE MFE and Subopts #1–9 were predicted using the SHAPE experimental data as constraints. Default MFE was predicted without the SHAPE data. Each secondary structure is shown as an arc diagram, in which the sequence is arranged along a horizontal line and base pairs are shown as arcs connecting the corresponding bases. The structures are listed in order of ascending pseudo-energy values. Pseudo-energy is the calculated free energy that includes the SHAPE pseudo-energy terms. Also shown are the energy values evaluated using the default energy model parameters ignoring SHAPE terms. MLD is the maximum ladder distance. All structures predicted using *RNAstructure* version 5.3.

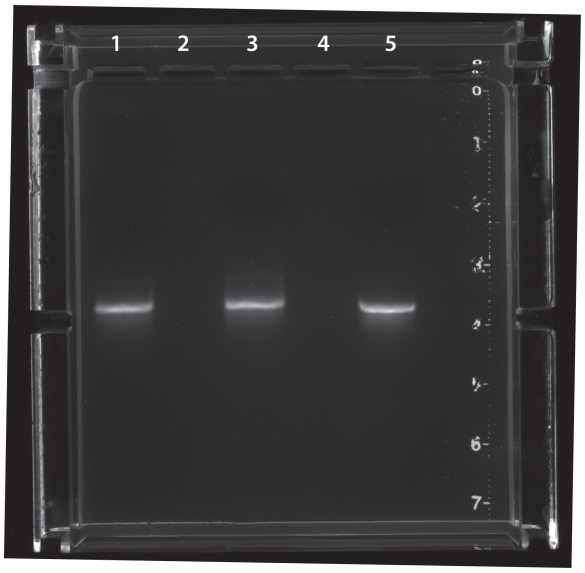


Figure 35: *In vitro* transcribed STMV RNA runs as a single band on a native gel. STMV RNA is run on a 1% agarose gel. No sample was loaded in lanes 2 or 4. Lanes 1 and 3 contain STMV RNA in SHAPE probing buffer without Mg^{2+} (50 mM HEPES pH 8.0, 200 mM sodium acetate pH 8.0) and lane 5 contains STMV RNA in 100 mM Tris-HCl pH 8.0. All samples were heated to 90°C for 2 min. Samples in lanes 1 and 5 were snap-cooled by chilling on ice, while the one in lane 3 was allowed to slow-cool to room temperature. The samples were then loaded on the gel using 6X native gel loading dye (New England Biolabs) and stained with SYBR Gold nucleic acid gel stain (Invitrogen). Lanes 1, 3 and 5 contain a single band, indicating a single dominant conformation.

REFERENCES

[1] Agalarov, S. C., Selivanova, O. M., Zheleznyakova, E. N., Zheleznaya, L. A., Matvienko, N. I., and Spirin, A. S., "Independent in vitro assembly of all three major morphological parts of the 30S ribosomal subunit of Thermus thermophilus," *Eur. J. Biochem.*, vol. 266, no. 2, pp. 533–537, 1999.

[2] Andrade, L. and Manolakos, E. S., "Signal background estimation and baseline correction algorithms for accurate DNA sequencing," *J. VLSI Signal Proc.*, vol. 35, no. 3, pp. 229–243, 2003.

[3] Athavale, S. S., Gossett, J. J., Bowman, J. C., Hud, N. V., Williams, L. D., and Harvey, S. C., "In vitro secondary structure of the genomic RNA of satellite tobacco mosaic virus," *PLoS ONE*, vol. 8, no. 1, p. e54384, 2013.

[4] Athavale, S. S., Gossett, J. J., Hsiao, C., Bowman, J. C., O'Neill, E., Hershkovitz, E., Preeprem, T., Hud, N. V., Wartell, R. M., Harvey, S. C., and Williams, L. D., "Domain III of the T. thermophilus 23S rRNA folds independently to a near-native state," *RNA*, vol. 18, no. 4, pp. 752–758, 2012.

[5] Athavale, S. S., Petrov, A. S., Hsiao, C., Watkins, D., Prickett, C. D., Gossett, J. J., Lie, L., Bowman, J. C., O'Neill, E., Bernier, C. R., Hud, N. V., Wartell, R. M., Harvey, S. C., and Williams, L. D., "RNA folding and catalysis mediated by iron (II)," *PLoS ONE*, vol. 7, no. 5, p. e38024, 2012.

[6] Aviran, S., Trapnell, C., Lucks, J. B., Mortimer, S. A., Luo, S., Schroth, G. P., Doudna, J. A., Arkin, A. P., and Pachter, L., "Modeling and automation of sequencing-based characterization of RNA structure," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 27, pp. 11069–11074, 2011.

[7] Badorrek, C. S. and Weeks, K. M., "Architecture of a gamma retroviral genomic RNA dimer," *Biochemistry*, vol. 45, no. 42, pp. 12664–12672, 2006.

[8] Bailey, K., Bettelheim, F. R., Lorand, L., and Middlebrook, W. R., "Action of thrombin in the clotting of fibrinogen," *Nature*, vol. 167, no. 4241, pp. 233–234, 1951.

[9] Baker, N. A., Sept, D., Joseph, S., Holst, M. J., and McCammon, J. A., "Electrostatics of nanosystems: Application to microtubules and the ribosome," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 18, pp. 10037–10041, 2001.

[10] Ban, N., Nissen, P., Hansen, J., Moore, P. B., and Steitz, T. A., "The complete atomic structure of the large ribosomal subunit at 2.4 angstrom resolution," *Science*, vol. 289, no. 5481, pp. 905–920, 2000.

[11] Belousoff, M. J., Davidovich, C., Zimmerman, E., Caspi, Y., Wekselman, I., Rozenszajn, L., Shapira, T., Sade-Falk, O., Taha, L., Bashan, A., Weiss,

M. S., and Yonath, A., "Ancient machinery embedded in the contemporary ribosome," *Biochem. Soc. Trans.*, vol. 38, pp. 422–427, 2010.

[12] Berno, A. J., "A graph theoretic approach to the analysis of DNA sequencing data," *Genome Res.*, vol. 6, no. 2, pp. 80–91, 1996.

[13] Bettelheim, F. R. and Bailey, K., "The products of the action of thrombin on fibrinogen," *Biochim. Biophys. Acta*, vol. 9, no. 5, pp. 578–579, 1952.

[14] Betts, L., Merenbloom, B. K., and Lord, S. T., "The structure of fibrinogen fragment D with the 'A' knob peptide GPRVVE," *J. Thromb. Haemost.*, vol. 4, no. 5, pp. 1139–1141, 2006.

[15] Bindewald, E., Wendeler, M., Legiewicz, M., Bona, M. K., Wang, Y., Pritt, M. J., Le Grice, S. F. J., and Shapiro, B. A., "Correlating SHAPE signatures with three-dimensional RNA structures," *RNA*, vol. 17, no. 9, pp. 1688–1696, 2011.

[16] Blomback, B., Blomback, M., Hessel, B., Iwanaga, S., and Reuterby, J., "Primary structure of human fibrinogen and fibrin. I. cleavage of fibrinogen with cyanogen bromide. isolation and characterization of NH2-terminal fragments of the alpha("A") chain," *J. Biol. Chem.*, vol. 247, no. 5, pp. 1496–1512, 1972.

[17] Bokov, K. and Steinberg, S. V., "A hierarchical model for evolution of 23S ribosomal RNA," *Nature*, vol. 457, no. 7232, pp. 977–980, 2009.

[18] Bowley, S. R., Merenbloom, B. K., Okumura, N., Betts, L., Heroux, A., Gorkun, O. V., and Lord, S. T., "Polymerization-defective fibrinogen variant gamma D364A binds knob "A" peptide mimic," *Biochemistry*, vol. 47, no. 33, pp. 8607–8613, 2008.

[19] Bowman, J. C., Lenz, T. K., Hud, N. V., and Williams, L. D., "Cations in charge: magnesium ions in RNA folding and catalysis," *Curr. Opin. Struct. Biol.*, vol. 22, no. 3, pp. 262–272, 2012.

[20] Brent, R. P., *Algorithms for minimization without derivatives.* Englewood Cliffs, N.J.: Prentice-Hall, 1972.

[21] Brimacombe, R., Maly, P., and Zwieb, C., "The structure of ribosomal RNA and its organization relative to ribosomal protein," *Prog. Nucleic Acid Res. Mol. Biol.*, vol. 28, pp. 1–48, 1983.

[22] Brion, P. and Westhof, E., "Hierarchy and dynamics of RNA folding," *Annu. Rev. Biophys. Biomol. Struct.*, vol. 26, pp. 113–137, 1997.

[23] Brosius, J., Dull, T. J., and Noller, H. F., "Complete nucleotide sequence of a 23S ribosomal RNA gene from Escherichia coli," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 77, no. 1, pp. 201–204, 1980.

[24] Campbell, I. D. and Downing, A. K., "Building protein structure and function from modular units," *Trends Biotechnol.*, vol. 12, no. 5, pp. 168–172, 1994.

[25] Cannone, J. J., Subramanian, S., Schnare, M. N., Collett, J. R., D'Souza, L. M., Du, Y. S., Feng, B., Lin, N., Madabusi, L. V., Muller, K. M., Pande, N., Shang, Z. D., Yu, N., and Gutell, R. R., "The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs," *BMC Bioinformatics*, vol. 3, p. 2, 2002.

[26] Canutescu, A. A. and Dunbrack, R. L., "Cyclic coordinate descent: A robotics algorithm for protein loop closure," *Protein Sci.*, vol. 12, no. 5, pp. 963–972, 2003.

[27] Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J., "The Amber biomolecular simulation programs," *J. Comput. Chem.*, vol. 26, no. 16, pp. 1668–1688, 2005.

[28] Chanhrasekaran, R. and Ramachandran, G. N., "Studies on the conformation of amino acids," *Int. J. Protein Res.*, vol. 2, no. 4, pp. 223–233, 1970.

[29] Chen, W., Guler, G., Kuruvilla, E., Schuster, G. B., Chiu, H.-C., and Riedo, E., "Development of self-organizing, self-directing molecular nanowires: Synthesis and characterization of conjoined DNA-2,5-bis(2-thienyl)pyrrole oligomers," *Macromolecules*, vol. 43, no. 9, pp. 4032–4040, 2010.

[30] Choi, Y. G., Dreher, T. W., and Rao, A. L. N., "tRNA elements mediate the assembly of an icosahedral RNA virus," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 2, pp. 655–660, 2002.

[31] Coleman, T. M., Wang, G., and Huang, F., "Superior 5' homogeneity of RNA from ATP-initiated transcription under the T7 phi 2.5 promoter," *Nucl. Acids Res.*, vol. 32, no. 1, p. e14, 2004.

[32] Cordero, P., Kladwang, W., VanLang, C. C., and Das, R., "Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference," *Biochemistry*, vol. 51, no. 36, pp. 7037–7039, 2012.

[33] Cornish-Bowden, A., "Detection of errors of interpretation in experiments in enzyme kinetics," *Methods*, vol. 24, no. 2, pp. 181–190, 2001.

[34] Crick, F. H. C., "Origin of genetic code," *J. Mol. Biol.*, vol. 38, no. 3, pp. 367–379, 1968.

[35] Darden, T., York, D., and Pedersen, L., "Particle mesh Ewald: An N.log(N) method for Ewald sums in large systems," *J. Chem. Phys.*, vol. 98, no. 12, pp. 10089–10092, 1993.

[36] Datta, B. and Schuster, G. B., "DNA-directed synthesis of aniline and 4-aminobiphenyl, oligomers: Programmed transfer of sequence information to a conjoined polymer nanowire," *J. Am. Chem. Soc.*, vol. 130, no. 10, pp. 2965–2973, 2008.

[37] Datta, B., Schuster, G. B., McCook, A., Harvey, S. C., and Zakrzewska, K., "DNA-directed assembly of polyanilines: Modified cytosine nucleotides transfer sequence programmability to a conjoined polymer," *J. Am. Chem. Soc.*, vol. 128, no. 45, pp. 14428–14429, 2006.

[38] Deigan, K. E., Li, T. W., Mathews, D. H., and Weeks, K. M., "Accurate SHAPE-directed RNA structure determination," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 1, pp. 97–102, 2009.

[39] Dodds, J. A., "Satellite tobacco mosaic virus," *Annu. Rev. Phytopathol.*, vol. 36, pp. 295–310, 1998.

[40] Doolittle, R. F., Chen, A., and Pandi, L., "Differences in binding specificity for the homologous gamma- and beta-chain "holes" on fibrinogen: exclusive binding of Ala-His-Arg-Pro-amide by the beta-chain hole," *Biochemistry*, vol. 45, no. 47, pp. 13962–13969, 2006.

[41] Doolittle, R. F. and Pandi, L., "Probing the beta-chain hole of fibrinogen with synthetic peptides that differ at their amino termini," *Biochemistry*, vol. 46, no. 35, pp. 10033–10038, 2007.

[42] Draper, D. E., "RNA folding: Thermodynamic and molecular descriptions of the roles of ions," *Biophys. J.*, vol. 95, pp. 5489–5495, 2008.

[43] Dreher, T. W., "Functions of the 3'-untranslated regions of positive strand RNA viral genomes," *Annu. Rev. Phytopathol.*, vol. 37, pp. 151–174, 1999.

[44] Egebjerg, J., Leffers, H., Christensen, A., Andersen, H., and Garrett, R. A., "Structure and accessibility of domain I of Escherichia coli 23S RNA in free RNA, in the L24-RNA complex and in 50S subunits - implications for ribosomal assembly," *J. Mol. Biol.*, vol. 196, no. 1, pp. 125–136, 1987.

[45] Eilers, P. H. C., "Parametric time warping," *Anal. Chem.*, vol. 76, no. 2, pp. 404–411, 2004.

[46] Felden, B., Florentz, C., McPherson, A., and Giege, R., "A histidine accepting tRNA-like fold at the 3'-end of satellite tobacco mosaic virus rna," *Nucl. Acids Res.*, vol. 22, no. 15, pp. 2882–2886, 1994.

[47] Fong, J. H., Geer, L. Y., Panchenko, A. R., and Bryant, S. H., "Modeling the evolution of protein domain architectures using maximum parsimony," *J. Mol. Biol.*, vol. 366, no. 1, pp. 307–315, 2007.

[48] Fox, G. E., "Origin and evolution of the ribosome," *Cold Spring Harbor Perspectives in Biology*, vol. 2, no. 9, p. a003483, 2010.

[49] Gallie, D. R., Feder, J. N., Schimke, R. T., and Walbot, V., "Functional analysis of the tobacco mosaic virus tRNA-like structure in cytoplasmic gene regulation," *Nucl. Acids Res.*, vol. 19, no. 18, pp. 5031–5036, 1991.

[50] Geer, C. B., Tripathy, A., Schoenfisch, M. H., Lord, S. T., and Gorkun, O. V., "Role of 'B-b' knob-hole interactions in fibrin binding to adsorbed fibrinogen," *J. Thromb. Haemost.*, vol. 5, no. 12, pp. 2344–2351, 2007.

[51] Gilbert, W., "Origin of life - the RNA world," *Nature*, vol. 319, no. 6055, pp. 618–618, 1986.

[52] Gossele, V., Fache, I., Meulewaeter, F., Cornelissen, M., and Metzlaff, M., "SVISS - a novel transient gene silencing system for gene function discovery and validation in tobacco plants," *Plant J.*, vol. 32, no. 5, pp. 859–866, 2002.

[53] Gossett, J. J. and Harvey, S. C., "Computational screening and design of DNA-linked molecular nanowires," *Nano Lett.*, vol. 11, no. 2, pp. 604–608, 2011.

[54] Guex, N. and Peitsch, M. C., "SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling," *Electrophoresis*, vol. 18, no. 15, pp. 2714–2723, 1997.

[55] Gultyaev, A. P., Van Batenburg, E., and Pleij, C. W. A., "Similarities between the secondary structure of satellite tobacco mosaic virus and tobamovirus RNAs," *J. Gen. Virol.*, vol. 75, pp. 2851–2856, 1994.

[56] Harms, J., Schluenzen, F., Zarivach, R., Bashan, A., Gat, S., Agmon, I., Bartels, H., Franceschi, F., and Yonath, A., "High resolution structure of the large ribosomal subunit from a mesophilic eubacterium," *Cell*, vol. 107, no. 5, pp. 679–688, 2001.

[57] Ho, B. K. and Dill, K. A., "Folding very short peptides using molecular dynamics," *PLoS Comput. Biol.*, vol. 2, no. 4, pp. 228–237, 2006.

[58] Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P., "Fast folding and comparison of RNA secondary structures," *Monatshefte Fur Chemie*, vol. 125, no. 2, pp. 167–188, 1994.

[59] Hsiao, C., Lenz, T. K., Peters, J. K., Fang, P.-Y., Schneider, D. M., Anderson, E. J., Preeprem, T., Bowman, J. C., O'Neill, E. B., Lie, L., Athavale, S. S., Gossett, J. J., Trippe, C., Murray, J., Petrov, A. S., Wartell, R. M., Harvey, S. C., Hud, N. V., and Dean Williams, L., "Molecular paleontology: a biochemical model of the ancestral ribosome," *Nucl. Acids Res.*, 2013.

[60] Hsiao, C., Mohan, S., Kalahar, B. K., and Williams, L. D., "Peeling the onion: Ribosomes are ancient molecular fossils," *Mol. Biol. Evol.*, vol. 26, no. 11, pp. 2415–2425, 2009.

[61] Hsiao, C. and Williams, L. D., "A recurrent magnesium-binding motif provides a framework for the ribosomal peptidyl transferase center," *Nucl. Acids Res.*, vol. 37, no. 10, pp. 3134–3142, 2009.

[62] Humphrey, W., Dalke, A., and Schulten, K., "VMD: Visual molecular dynamics," *J. Mol. Graphics Modell.*, vol. 14, no. 1, pp. 33–38, 1996.

[63] Hury, J., Nagaswamy, U., Larios-Sanz, M., and Fox, G. E., "Ribosome origins: The relative age of 23S rRNA domains," *Orig. Life Evol. Biosph.*, vol. 36, no. 4, pp. 421–429, 2006.

[64] Hyeon, C., Dima, R. I., and Thirumalai, D., "Size, shape, and flexibility of RNA structures," *J. Chem. Phys.*, vol. 125, no. 19, p. 194905, 2006.

[65] Karabiber, F., McGinnis, J. L., Favorov, O. V., and Weeks, K. M., "QuShape: Rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis," *RNA*, vol. 19, pp. 63–73, 2013.

[66] Kenyon, J. C., Tanner, S. J., Legiewicz, M., Phillip, P. S., Rizvi, T. A., Le Grice, S. F. J., and Lever, A. M. L., "Shape analysis of the fiv leader rna reveals a structural switch potentially controlling viral packaging and genome dimerization," *Nucl. Acids Res.*, vol. 39, no. 15, pp. 6692–6704, 2011.

[67] Kladwang, W., VanLang, C. C., Cordero, P., and Das, R., "Understanding the errors of SHAPE-directed RNA structure modeling," *Biochemistry*, vol. 50, no. 37, pp. 8049–8056, 2011.

[68] Kostelansky, M. S., Betts, L., Gorkun, O. V., and Lord, S. T., "2.8 angstrom crystal structures of recombinant fibrinogen fragment D with and without two peptide ligands: GHRP binding to the "b" site disrupts its nearby calcium-binding site," *Biochemistry*, vol. 41, no. 40, pp. 12124–12132, 2002.

[69] Kuznetsov, Y. G., Dowell, J. J., Gavira, J. A., Ng, J. D., and McPherson, A., "Biophysical and atomic force microscopy characterization of the RNA from satellite tobacco mosaic virus," *Nucl. Acids Res.*, vol. 38, no. 22, pp. 8284–8294, 2010.

[70] Larson, S. B., Day, J., Greenwood, A., and McPherson, A., "Refined structure of satellite tobacco mosaic virus at 1.8 angstrom resolution," *J. Mol. Biol.*, vol. 277, no. 1, pp. 37–59, 1998.

[71] Larson, S. B. and McPherson, A., "Satellite tobacco mosaic virus RNA: Structure and implications for assembly," *Curr. Opin. Struct. Biol.*, vol. 11, pp. 59–65, 2001.

[72] Laudano, A. P. and Doolittle, R. F., "Synthetic peptide derivatives that bind to fibrinogen and prevent the polymerization of fibrin monomers," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 75, no. 7, pp. 3085–3089, 1978.

[73] Laudano, A. P. and Doolittle, R. F., "Studies on synthetic peptides that bind to fibrinogen and prevent fibrin polymerization. structural requirements, number of binding sites, and species differences," *Biochemistry*, vol. 19, no. 5, pp. 1013–1019, 1980.

[74] Laudano, A. P. and Doolittle, R. F., "Influence of calcium ion on the binding of fibrin amino terminal peptides to fibrinogen," *Science*, vol. 212, no. 4493, pp. 457–459, 1981.

[75] Laurent, T. C. and Blomback, B., "On the significance of the release of 2 different peptides from fibrinogen during clotting," *Acta Chem. Scand.*, vol. 12, no. 9, pp. 1875–1877, 1958.

[76] Leach, A. R., *Molecular Modelling: Principles and Applications*. Prentice Hall, 2nd ed., 2001.

[77] Leffers, H., Egebjerg, J., Andersen, A., Christensen, T., and Garrett, R. A., "Domain VI of Escherichia coli 23S ribosomal RNA - structure, assembly and function," *J. Mol. Biol.*, vol. 204, no. 3, pp. 507–522, 1988.

[78] Leontis, N. B., Stombaugh, J., and Westhof, E., "The non-Watson-Crick base pairs and their associated isostericity matrices," *Nucl. Acids Res.*, vol. 30, no. 16, pp. 3497–3531, 2002.

[79] Leontis, N. B. and Westhof, E., "Geometric nomenclature and classification of RNA base pairs," *RNA*, vol. 7, no. 4, pp. 499–512, 2001.

[80] Lexa, K. W., Alser, K. A., Salisburg, A. M., Ellens, D. J., Hernandez, L., Bono, S. J., Michael, H. C., Derby, J. R., Skiba, J. G., Feldgus, S., Kirschner, K. N., and Shields, G. C., "The search for low energy conformational families of small peptides: Searching for active conformations of small peptides in the absence of a known receptor," *Int. J. Quantum Chem.*, vol. 107, no. 15, pp. 3001–3012, 2007.

[81] Litvinov, R. I., Gorkun, O. V., Owen, S. F., Shuman, H., and Weisel, J. W., "Polymerization of fibrin: specificity, strength, and stability of knob-hole interactions studied at the single-molecule level," *Blood*, vol. 106, no. 9, pp. 2944–2951, 2005.

[82] Litvinov, R. I., Gorkun, O. V., Galanakis, D. K., Yakovlev, S., Medved, L., Shuman, H., and Weisel, J. W., "Polymerization of fibrin: direct observation and quantification of individual B:b knob-hole interactions," *Blood*, vol. 109, no. 1, pp. 130–138, 2007.

[83] Liu, Y., Wang, R., Ding, L., Sha, R., Lukeman, P. S., Canary, J. W., and Seeman, N. C., "Thermodynamic analysis of nylon nucleic acids," *Chembiochem*, vol. 9, no. 10, pp. 1641–1648, 2008.

[84] Lorand, L. and Middlebrook, W. R., "The action of thrombin on fibrinogen," *Biochem. J.*, vol. 52, no. 2, pp. 196–199, 1952.

[85] Lorand, L. and Middlebrook, W. R., "Studies on fibrino-peptide," *Biochim. Biophys. Acta*, vol. 9, no. 5, pp. 581–582, 1952.

[86] Low, J. T. and Weeks, K. M., "SHAPE-directed RNA secondary structure prediction," *Methods*, vol. 52, no. 2, pp. 150–158, 2010.

[87] Lyman, E. and Zuckerman, D. M., "Ensemble-based convergence analysis of biomolecular trajectories," *Biophys. J.*, vol. 91, no. 1, pp. 164–172, 2006.

[88] MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D., and Karplus, M., "All-atom empirical potential for molecular modeling and dynamics studies of proteins," *J. Phys. Chem. B*, vol. 102, no. 18, pp. 3586–3616, 1998.

[89] Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., and Turner, D. H., "Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 19, pp. 7287–7292, 2004.

[90] Matthews, B. W., Nicholson, H., and Becktel, W. J., "Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 84, no. 19, pp. 6663–6667, 1987.

[91] Mayrand, P. E., Corcoran, K. P., Ziegle, J. S., Robertson, J. M., Hoff, L. B., and Kronick, M. N., "The use of fluorescence detection and internal lane standards to size PCR products automatically," *Appl. Theor. Electrophor.*, vol. 3, no. 1, pp. 1–11, 1992.

[92] McGinnis, J. L., Dunkle, J. A., Cate, J. H. D., and Weeks, K. M., "The mechanisms of RNA SHAPE chemistry," *J. Am. Chem. Soc.*, vol. 134, no. 15, pp. 6617–6624, 2012.

[93] Merino, E. J., Wilkinson, K. A., Coughlan, J. L., and Weeks, K. M., "RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE)," *J. Am. Chem. Soc.*, vol. 127, no. 12, pp. 4223–4231, 2005.

[94] Mills, J. E. J. and Dean, P. M., "Three-dimensional hydrogen-bond geometry and probability information from a crystal survey," *J. Comput. Aided Mol. Des.*, vol. 10, no. 6, pp. 607–622, 1996.

[95] Mirkov, T. E., Kurath, G., Mathews, D. M., Elliott, K., Dodds, J. A., and Fitzmaurice, L., "Factors affecting efficient infection of tobacco with in vitro RNA transcripts from cloned cDNAs of satellite tobacco mosaic virus," *Virology*, vol. 179, no. 1, pp. 395–402, 1990.

[96] Mirkov, T. E., Mathews, D. M., Du Plessis, D. H., and Dodds, J. A., "Nucleotide sequence and translation of satellite tobacco mosaic virus RNA," *Virology*, vol. 170, no. 1, pp. 139–146, 1989.

[97] Mitra, S., Shcherbakova, I. V., Altman, R. B., Brenowitz, M., and Laederach, A., "High-throughput single-nucleotide structural mapping by capillary automated footprinting analysis," *Nucl. Acids Res.*, vol. 36, no. 11, p. e63, 2008.

[98] Mortimer, S. A. and Weeks, K. M., "A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry," *J. Am. Chem. Soc.*, vol. 129, no. 14, pp. 4144–4145, 2007.

[99] Mortimer, S. A. and Weeks, K. M., "Time-resolved RNA SHAPE chemistry," *J. Am. Chem. Soc.*, vol. 130, no. 48, pp. 16178–16180, 2008.

[100] Morton, T. A. and Myszka, D. G., "Kinetic analysis of macromolecular interactions using surface plasmon resonance biosensors," *Methods Enzymol.*, vol. 295, pp. 268–294, 1998.

[101] Morton, T. A., Myszka, D. G., and Chaiken, I. M., "Interpreting complex binding-kinetics from optical biosensors - a comparison of analysis by linearization, the integrated rate-equation, and numerical-integration," *Anal. Biochem.*, vol. 227, no. 1, pp. 176–185, 1995.

[102] Myszka, D. G. and Morton, T. A., "CLAMP: A biosensor kinetic data analysis program," *Trends Biochem. Sci.*, vol. 23, no. 4, pp. 149–150, 1998.

[103] Myszyka, D. G., "Improving biosensor analysis," *J. Mol. Recognit.*, vol. 12, no. 5, pp. 279–284, 1999.

[104] Nissen, P., Hansen, J., Ban, N., Moore, P. B., and Steitz, T. A., "The structural basis of ribosome activity in peptide bond synthesis," *Science*, vol. 289, no. 5481, pp. 920–930, 2000.

[105] Noller, H. F., "RNA structure: Reading the ribosome," *Science*, vol. 309, no. 5740, pp. 1508–1514, 2005.

[106] Noller, H. F., Hoffarth, V., and Zimniak, L., "Unusual resistance of peptidyl transferase to protein extraction procedures," *Science*, vol. 256, no. 5062, pp. 1416–1419, 1992.

[107] Noller, H. F., Kop, J., Wheaton, V., Brosius, J., Gutell, R. R., Kopylov, A. M., Dohme, F., Herr, W., Stahl, D. A., Gupta, R., and Woese, C. R., "Secondary structure model for 23S ribosomal RNA," *Nucl. Acids Res.*, vol. 9, no. 22, pp. 6167–6189, 1981.

[108] Ogle, J. M., Brodersen, D. E., Clemons, W. M., Tarry, M. J., Carter, A. P., and Ramakrishnan, V., "Recognition of cognate transfer RNA by the 30S ribosomal subunit," *Science*, vol. 292, no. 5518, pp. 897–902, 2001.

[109] Orgel, L. E., "Evolution of genetic apparatus," *J. Mol. Biol.*, vol. 38, no. 3, pp. 381–393, 1968.

[110] Ostergaard, P., Phan, H., Johansen, L. B., Egebjerg, J., Ostergaard, L., Porse, B. T., and Garrett, R. A., "Assembly of proteins and 5 S rRNA to transcripts of the major structural domains of 23 S rRNA," *J. Mol. Biol.*, vol. 284, no. 2, pp. 227–240, 1998.

[111] Pang, P. S., Elazar, M., Pham, E. A., and Glenn, J. S., "Simplified rna secondary structure mapping by automation of shape data analysis," *Nucl. Acids Res.*, vol. 39, no. 22, p. e151, 2011.

[112] Parsons, J., Holmes, J. B., Rojas, J. M., Tsai, J., and Strauss, C. E. M., "Practical conversion from torsion space to Cartesian space for in silico protein synthesis," *J. Comput. Chem.*, vol. 26, no. 10, pp. 1063–1068, 2005.

[113] Petrie, T. A., Capadona, J. R., Reyes, C. D., and Garcia, A. J., "Integrin specificity and enhanced cellular activities associated with surfaces presenting a recombinant fibronectin fragment compared to RGD supports," *Biomaterials*, vol. 27, no. 31, pp. 5459–5470, 2006.

[114] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E., "UCSF chimera - a visualization system for exploratory research and analysis," *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, 2004.

[115] Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L., and Schulten, K., "Scalable molecular dynamics with NAMD," *J. Comput. Chem.*, vol. 26, no. 16, pp. 1781–1802, 2005.

[116] Ponder, J. W. and Richards, F. M., "Tertiary templates for proteins - use of packing criteria in the enumeration of allowed sequences for different structural classes," *J. Mol. Biol.*, vol. 193, no. 4, pp. 775–791, 1987.

[117] Rao, A. L. N., Dreher, T. W., Marsh, L. E., and Hall, T. C., "Telomeric function of the tRNA-like structure of brome mosaic virus RNA," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 86, no. 14, pp. 5335–5339, 1989.

[118] Reuter, J. S. and Mathews, D. H., "RNAstructure: software for RNA secondary structure prediction and analysis," *BMC Bioinformatics*, vol. 11, p. 129, 2010.

[119] Rich, A., *On the problems of evolution and biochemical information transfer*. New York: Academic Press, 1962. in Horizons in Biochemistry.

[120] Routh, G., Dodds, J. A., Fitzmaurice, L., and Mirkov, T. E., "Characterization of deletion and frameshift mutants of satellite tobacco mosaic virus," *Virology*, vol. 212, no. 1, pp. 121–127, 1995.

[121] Ryckaert, J. P., Ciccotti, G., and Berendsen, H. J. C., "Numerical-integration of cartesian equations of motion of a system with constraints - molecular-dynamics of n-alkanes," *J. Comput. Phys.*, vol. 23, no. 3, pp. 327–341, 1977.

[122] Saenger, W., *Principles of nucleic acid structure*. New York: Springer-Verlag, 1984.

[123] Samaha, R. R., Obrien, B., Obrien, T. W., and Noller, H. F., "Independent in vitro assembly of a ribonucleoprotein particle containing the 3' domain of 16S rRNA," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 91, no. 17, pp. 7884–7888, 1994.

[124] Sampson, J. R. and Uhlenbeck, O. C., "Biochemical and physical characterization of an unmodified yeast phenylalanine transfer RNA transcribed in vitro," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 85, no. 4, pp. 1033–1037, 1988.

[125] Schimmel, P. R. and Flory, P. J., "Conformational energies and configurational statistics of copolypeptides containing L-proline," *J. Mol. Biol.*, vol. 34, no. 1, pp. 105–120, 1968.

[126] Schlick, T., *Molecular Modeling and Simulation*. Springer, 2002.

[127] Schneemann, A., "The structural and functional role of RNA in icosahedral virus assembly," *Annu. Rev. Microbiol.*, vol. 60, pp. 51–67, 2006.

[128] Schroeder, S. J., Stone, J. W., Bleckley, S., Gibbons, T., and Mathews, D. M., "Ensemble of secondary structures for encapsidated satellite tobacco mosaic virus RNA consistent with chemical probing and crystallography constraints," *Biophys. J.*, vol. 101, no. 1, pp. 167–75, 2011.

[129] Schuwirth, B. S., Borovinskaya, M. A., Hau, C. W., Zhang, W., Vila-Sanjurjo, A., Holton, J. M., and Cate, J. H. D., "Structures of the bacterial ribosome at 3.5 angstrom resolution," *Science*, vol. 310, no. 5749, pp. 827–834, 2005.

[130] Selmer, M., Dunham, C. M., Murphy, F. V., Weixlbaumer, A., Petry, S., Kelley, A. C., Weir, J. R., and Ramakrishnan, V., "Structure of the 70S ribosome complexed with mRNA and tRNA," *Science*, vol. 313, no. 5795, pp. 1935–1942, 2006.

[131] Sloof, P., Vandenburg, J., Voogd, A., Benne, R., Agostinelli, M., Borst, P., Gutell, R., and Noller, H., "Further characterization of the extremely small mitochondrial ribosomal RNAs from trypanosomes: a detailed comparison of the 9S and 12S RNAs from Crithidia fasciculata and Trypanosoma brucei with rRNAs from other organisms," *Nucl. Acids Res.*, vol. 13, no. 11, pp. 4171–4190, 1985.

[132] Smith, T. F., Lee, J. C., Gutell, R. R., and Hartman, H., "The origin and evolution of the ribosome," *Biology Direct*, vol. 3, p. 16, 2008.

[133] Soll, D. and RajBhandary, U., *tRNA: structure, biosynthesis, and function.* Washington, D.C.: ASM Press, 1995.

[134] Spraggon, G., Everse, S. J., and Doolittle, R. F., "Crystal structures of fragment D from human fibrinogen and its crosslinked counterpart from fibrin," *Nature*, vol. 389, no. 6650, pp. 455–462, 1997.

[135] Srinivasan, S. and Schuster, G. B., "A conjoined thienopyrrole oligomer formed by using DNA as a molecular guide," *Org. Lett.*, vol. 10, no. 17, pp. 3657–3660, 2008.

[136] Stabenfeldt, S. E., Gossett, J. J., and Barker, T. H., "Building better fibrin knob mimics: an investigation of synthetic fibrin knob peptide structures in solution and their dynamic binding with fibrinogen/fibrin holes," *Blood*, vol. 116, no. 8, pp. 1352–1359, 2010.

[137] Sugeta, H. and Miyazawa, T., "General method for calculating helical parameters of polymer chains from bond lengths bond angles and internal-rotation angles," *Biopolymers*, vol. 5, no. 7, pp. 673–679, 1967.

[138] Swenson, M. S., Anderson, J., Ash, A., Gaurav, P., Sukosd, Z., Bader, D. A., Harvey, S. C., and Heitsch, C. E., "GTfold: Enabling parallel RNA secondary structure prediction on multi-core desktops," *BMC Res. Notes*, vol. 5, no. 1, p. 341, 2012.

[139] Tumminia, S. J., Hellmann, W., Wall, J. S., and Boublik, M., "Visualization of protein-nucleic acid interactions involved in the in vitro assembly of the Escherichia coli 50 S ribosomal subunit," *J. Mol. Biol.*, vol. 235, no. 4, pp. 1239–1250, 1994.

[140] Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R., and Leunissen, J. A., "Primer3Plus, an enhanced web interface to Primer3," *Nucl. Acids Res.*, vol. 35, no. Web Server issue, pp. W71–74, 2007.

[141] Vanhoof, G., Goossens, F., Demeester, I., Hendriks, D., and Scharpe, S., "Proline motifs in peptides and their biological processing," *FASEB J.*, vol. 9, no. 9, pp. 736–744, 1995.

[142] Vasa, S. M., Guex, N., Wilkinson, K. A., Weeks, K. M., and Giddings, M. C., "ShapeFinder: A software system for high-throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis," *RNA*, vol. 14, pp. 1979–1990, 2008.

[143] Vicens, Q., Gooding, A. R., Laederach, A., and Cech, T. R., "Local RNA structural changes induced by crystallization are revealed by SHAPE," *RNA*, vol. 13, no. 4, pp. 536–548, 2007.

[144] Wang, B., Wilkinson, K. A., and Weeks, K. M., "Complex ligand-induced conformational changes in trna(asp) revealed by single-nucleotide resolution shape chemistry," *Biochemistry*, vol. 47, no. 11, pp. 3454–3461, 2008.

[145] Wang, J. M., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A., "Development and testing of a general Amber force field," *J. Comput. Chem.*, vol. 25, no. 9, pp. 1157–1174, 2004.

[146] Wang, L. C. T. and Chen, C. C., "A combined optimization method for solving the inverse kinematics problem of mechanical manipulators," *IEEE Transactions on Robotics and Automation*, vol. 7, no. 4, pp. 489–499, 1991.

[147] Wang, M. B., Bian, X. Y., Wu, L. M., Liu, L. X., Smith, N. A., Isenegger, D., Wu, R. M., Masuta, C., Vance, V. B., Watson, J. M., Rezaian, A., Dennis, E. S., and Waterhouse, P. M., "On the role of RNA silencing in the pathogenicity and evolution of viroids and viral satellites," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 9, pp. 3275–3280, 2004.

[148] Watson, J. D. and Crick, F. H., "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid," *Nature*, vol. 171, no. 4356, pp. 737–738, 1953.

[149] Watts, J. M., Dang, K. K., Gorelick, R. J., Leonard, C. W., Bess, J. W., Swanstrom, R., Burch, C. L., and Weeks, K. M., "Architecture and secondary structure of an entire HIV-1 RNA genome," *Nature*, vol. 460, no. 7256, pp. 711–716, 2009.

[150] Weeks, K. M., "Advances in rna structure analysis by chemical probing," *Curr. Opin. Struct. Biol.*, vol. 20, no. 3, pp. 295–304, 2010.

[151] Weeks, K. M. and Mauger, D. M., "Exploring rna structural codes with shape chemistry," *Acc. Chem. Res.*, vol. 44, no. 12, pp. 1280–1291, 2011.

[152] Weitzmann, C. J., Cunningham, P. R., Nurse, K., and Ofengand, J., "Chemical evidence for domain assembly of the Escherichia coli 30S ribosome," *FASEB J.*, vol. 7, no. 1, pp. 177–180, 1993.

[153] Wilkinson, K. A., Gorelick, R. J., Vasa, S. M., Guex, N., Rein, A., Mathews, D. H., Giddings, M. C., and Weeks, K. M., "High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states," *PLoS Biol.*, vol. 6, no. 4, pp. 883–999, 2008.

[154] Wilkinson, K. A., Merino, E. J., and Weeks, K. M., "RNA SHAPE chemistry reveals nonhierarchical interactions dominate equilibrium structural transitions in tRNA(asp) transcripts," *J. Am. Chem. Soc.*, vol. 127, no. 13, pp. 4659–4667, 2005.

[155] Wilkinson, K. A., Merino, E. J., and Weeks, K. M., "Selective 2 '-hydroxyl acylation analyzed by primer extension (shape): quantitative rna structure analysis at single nucleotide resolution," *Nature Protocols*, vol. 1, no. 3, pp. 1610–1616, 2006.

[156] Wilkinson, K. A., Vasa, S. M., Deigan, K. E., Mortimer, S. A., Giddings, M. C., and Weeks, K. M., "Influence of nucleotide identity on ribose 2'-hydroxyl reactivity in RNA," *RNA*, vol. 15, no. 7, pp. 1314–1321, 2009.

[157] Wimberly, B. T., Brodersen, D. E., Clemons, W. M., Morgan-Warren, R. J., Carter, A. P., Vonrhein, C., Hartsch, T., and Ramakrishnan, V., "Structure of the 30S ribosomal subunit," *Nature*, vol. 407, no. 6802, pp. 327–339, 2000.

[158] Wirapati, P. J., *An Automated Allele-calling System for High-throughput Microsatellite Genotyping*. PhD thesis, University of Melbourne, 2003.

[159] Woese, C. R., *The genetic code: The molecular basis for genetic expression*. New York: Harper & Row, 1967.

[160] Woese, C. R., "Translation: In retrospect and prospect," *RNA*, vol. 7, no. 8, pp. 1055–1067, 2001.

[161] Woese, C. R., Magrum, L. J., Gupta, R., Siegel, R. B., Stahl, D. A., Kop, J., Crawford, N., Brosius, J., Gutell, R., Hogan, J. J., and Noller, H. F., "Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence," *Nucl. Acids Res.*, vol. 8, no. 10, pp. 2275–2293, 1980.

[162] Wolf, Y. I. and Koonin, E. V., "On the origin of the translation system and the genetic code in the RNA world by means of natural selection, exaptation, and subfunctionalization," *Biology Direct*, vol. 2, p. 14, 2007.

[163] Xu, W., Bolduc, F., Hong, N., and Perreault, J.-P., "The use of a combination of computer-assisted structure prediction and SHAPE probing to elucidate the secondary structures of five viroids," *Mol. Plant. Pathol.*, vol. 13, no. 7, pp. 666–676, 2012.

[164] Yang, H. W., Jossinet, F., Leontis, N., Chen, L., Westbrook, J., Berman, H., and Westhof, E., "Tools for the automatic identification and classification of RNA base pairs," *Nucl. Acids Res.*, vol. 31, no. 13, pp. 3450–3460, 2003.

[165] Yoffe, A. M., Prinsen, P., Gopal, A., Knobler, C. M., Gelbart, W. M., and Ben-Shaul, A., "Predicting the sizes of large RNA molecules," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, pp. 16153–16158, 2008.

[166] Yoon, S., Kim, J., Hum, J., Kim, H., Park, S., Kladwang, W., and Das, R., "HiTRACE: high-throughput robust analysis for capillary electrophoresis," *Bioinformatics*, vol. 27, no. 13, pp. 1798–1805, 2011.

[167] Yusupov, M. M., Yusupova, G. Z., Baucom, A., Lieberman, K., Earnest, T. N., Cate, J. H. D., and Noller, H. F., "Crystal structure of the ribosome at 5.5 angstrom resolution," *Science*, vol. 292, no. 5518, pp. 883–896, 2001.

[168] Zeng, Y., Larson, S. B., Heitsch, C. E., McPherson, A., and Harvey, S. C., "A model for the structure of satellite tobacco mosaic virus," *J. Struct. Biol.*, vol. 180, no. 1, pp. 110–116, 2012.

[169] Zheng, G., Lu, X.-J., and Olson, W. K., "Web 3DNA—a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures," *Nucl. Acids Res.*, vol. 37, pp. W240–W246, 2009.

[170] Zhu, L., Lukeman, P. S., Canary, J. W., and Seeman, N. C., "Nylon/DNA: Single-stranded DNA with a covalently stitched nylon lining," *J. Am. Chem. Soc.*, vol. 125, no. 34, pp. 10178–10179, 2003.

[171] Ziegle, J. S., Su, Y., Corcoran, K. P., Nie, L., Mayrand, P. E., Hoff, L. B., Mcbride, L. J., Kronick, M. N., and Diehl, S. R., "Application of automated DNA sizing technology for genotyping microsatellite loci," *Genomics*, vol. 14, no. 4, pp. 1026–1031, 1992.

[172] Zuker, M., "On finding all suboptimal foldings of an RNA molecule," *Science*, vol. 244, no. 4900, pp. 48–52, 1989.

[173] Zuker, M. and Sankoff, D., "RNA secondary structures and their prediction," *Bull. Math. Biol.*, vol. 46, no. 4, pp. 591–621, 1984.